



PQStat Software
Statistical Computational Software

User Guide - PQStat

Barbara Wieckowska

COPYRIGHT ©2010-2014 PQSTAT SOFTWARE

All rights reserved

Version 1.4.8
P7909121213

www.pqstat.pl

Contents

1	SYSTEM REQUIREMENTS	5
2	HOW TO INSTALL	5
3	WORKING WITH DOCUMENTS	6
3.1	HOW TO WORK WITH DATASHEETS	8
3.1.1	HOW TO ADD, TO DELETE AND TO EXPORT DATASHEETS	8
3.1.2	HOW TO INSERT DATA INTO A SHEET	8
3.1.3	DATASHEET WINDOW	10
3.1.4	CELLS FORMAT	11
3.1.5	DATA EDITING	13
3.1.6	HOW TO SORT DATA	14
3.1.7	HOW TO CONVERT RAW DATA INTO CONTINGENCY TABLE	15
3.1.8	HOW TO CONVERT CONTINGENCY TABLE INTO RAW DATA	16
3.1.9	FORMULAS	16
3.1.10	HOW TO GENERATE DATA	20
3.1.11	MISSING DATA	21
3.1.12	NORMALIZATION/STANDARDIZATION	24
3.1.13	SIMILARITY MATRIX	25
3.2	HOW TO WORK WITH REPORTS (RESULTS SHEETS)	35
3.3	HOW TO CHANGE LANGUAGE SETTINGS IN PQSTAT?	36
3.4	MENU	37
4	HOW TO ORGANISE WORK WITH PQSTAT	41
4.1	HOW TO ORGANISE DATA	41
4.2	HOW TO REDUCE A DATASHEET WORKSPACE	43
4.3	MULTIPLE REPEATED ANALYSIS	47
4.4	INFORMATION GIVEN IN A REPORT	47
4.5	MARKING OF STATISTICALLY SIGNIFICANT RESULTS	47
5	GRAPHS	48
5.1	GRAPHS GALLERY	48
5.1.1	Bar plots	48
5.1.2	Error plots	53
5.1.3	Box-Whiskers plots	55
5.1.4	Scatter plots	56
5.1.5	Line plots	58
6	FREQUENCY TABLES AND EMPIRICAL DATA DISTRIBUTION	60
7	DESCRIPTIVE STATISTICS	65
7.1	MEASUREMENT SCALES	65
7.2	MEASURES OF POSITION (LOCATION)	67
7.2.1	CENTRAL TENDENCY MEASURES	67
7.2.2	ANOTHER MEASURES OF POSITION	68
7.3	MEASURES OF VARIABILITY (DISPERSION)	69
7.4	ANOTHER DISTRIBUTION CHARACTERISTICS	70
8	PROBABILITY DISTRIBUTIONS	73
8.1	CONTINUOUS PROBABILITY DISTRIBUTIONS	75
8.2	PROBABILITY DISTRIBUTION CALCULATOR	78
9	HYPOTHESES TESTING	81
9.0.1	POINT AND INTERVAL ESTIMATION	81
9.0.2	VERIFICATION OF STATISTICAL HYPOTHESES	81

10 COMPARISON - 1 GROUP	84
10.1 PARAMETRIC TESTS	85
10.1.1 The t -test for a single sample	85
10.2 NONPARAMETRIC TESTS	88
10.2.1 The Kolmogorov-Smirnov test and the Lilliefors test	88
10.2.2 The Wilcoxon test (signed-ranks)	91
10.2.3 The Chi-square goodness-of-fit test	94
10.2.4 Tests for proportion	97
11 COMPARISON - 2 GROUPS	101
11.1 PARAMETRIC TESTS	102
11.1.1 The Fisher-Snedecor test	102
11.1.2 The t -test for independent groups	103
11.1.3 The t -test with the Cochran-Cox adjustment	104
11.1.4 The t -test for dependent groups	107
11.2 NONPARAMETRIC TESTS	109
11.2.1 The Mann-Whitney U test	109
11.2.2 The Wilcoxon test (matched-pairs)	112
11.2.3 TESTS FOR CONTINGENCY TABLES	114
11.2.4 The Chi-square test for trend for $R \times 2$ tables	118
11.2.5 The Chi-square test and Fisher test for $R \times C$ tables	120
11.2.6 The Chi-square test and the Fisher test for 2×2 tables (with corrections)	125
11.2.7 Relative Risk and Odds Ratio	131
11.2.8 The Z test for 2 independent proportions	133
11.2.9 The McNemar test, the Bowker test of internal symmetry	136
11.2.10 Z Test for two dependent proportions	141
12 COMPARISON - MORE THAN 2 GROUPS	144
12.1 PARAMETRIC TESTS	145
12.1.1 The ANOVA for independent groups	145
12.1.2 The contrasts and the POST-HOC tests	146
12.1.3 The Brown-Forsythe test and the Levene test	151
12.1.4 The ANOVA for dependent groups	152
12.2 NONPARAMETRIC TESTS	156
12.2.1 The Kruskal-Wallis ANOVA	156
12.2.2 The Friedman ANOVA	158
12.2.3 The Chi-square test for multidimensional contingency tables	161
12.2.4 The Q-Cochran ANOVA	163
13 STRATIFIED ANALYSIS	167
13.1 THE MANTEL - HAENSZEL METHOD FOR SEVERAL 2×2 TABLES	167
13.1.1 The Mantel-Haenszel odds ratio	167
13.1.2 The Mantel-Haenszel relative risk	172
14 CORRELATION	174
14.1 PARAMETRIC TESTS	175
14.1.1 THE LINEAR CORRELATION COEFFICIENTS	175
14.1.2 The test of significance for the Pearson product-moment correlation coefficient	176
14.1.3 The test of significance for the coefficient of linear regression equation	176
14.1.4 The test for checking the equality of the Pearson product-moment correlation coefficients, which come from 2 independent populations	180
14.1.5 The test for checking the equality of the coefficients of linear regression equation, which come from 2 independent populations	181
14.2 NONPARAMETRIC TESTS	183
14.2.1 THE MONOTONIC CORRELATION COEFFICIENTS	183
14.2.2 The test of significance for the Spearman's rank-order correlation coefficient	184
14.2.3 The test of significance for the Kendall's tau correlation coefficient	186

14.2.4	CONTINGENCY TABLES COEFFICIENTS AND THEIR STATISTICAL SIGNIFICANCE	188
15	AGREEMENT ANALYSIS	194
15.1	PARAMETRIC TESTS	195
15.1.1	The intraclass correlation coefficient and the test of its significance	195
15.2	NONPARAMETRIC TESTS	199
15.2.1	The Kendall's coefficient of concordance and the test of its significance	199
15.2.2	The Cohen's Kappa coefficient and the test of its significance	202
16	DIAGNOSTIC TESTS	206
16.1	EVALUATION OF DIAGNOSTIC TEST	206
16.2	ROC CURVE	210
16.2.1	Selection of optimum cut-off	213
16.2.2	ROC curves comparison	217
17	MULTIDIMENSIONAL MODELS	224
17.1	PREPARATION OF THE VARIABLES FOR THE ANALYSIS IN MULTIDIMENSIONAL MODELS	224
17.1.1	Variable coding in multidimensional models	224
17.1.2	Interactions	227
17.2	MULTIPLE LINEAR REGRESSION	227
17.2.1	Model verification	229
17.2.2	More information about the variables in the model	231
17.2.3	Analysis of model residuals	232
17.2.4	Prediction on the basis of the model	233
17.3	COMPARISON OF MULTIPLE LINEAR REGRESSION MODELS	240
17.4	LOGISTIC REGRESSION	244
17.4.1	Odds Ratio	246
17.4.2	Model verification	247
17.5	COMPARISON OF LOGISTIC REGRESSION MODELS	260
18	DIMENSION REDUCTION AND GROUPING	264
18.1	PRINCIPAL COMPONENT ANALYSIS	264
18.1.1	The interpretation of coefficients related to the analysis	265
18.1.2	Graphical interpretation	266
18.1.3	The criteria of dimension reduction	268
18.1.4	Defining principal components	268
18.1.5	The advisability of using the Principal component analysis	269
19	SURVIVAL ANALYSIS	276
19.1	LIFE TABLES	277
19.2	KAPLAN-MEIER CURVES	280
19.3	COMPARISON OF SURVIVAL CURVES	282
19.3.1	Differences among the survival curves	284
19.3.2	Survival curve trend	285
19.3.3	Survival curves for the stratas	285
19.4	PROPORTIONAL COX HAZARD REGRESSION	292
19.4.1	Hazard ratio	294
19.4.2	Model verification	294
19.4.3	Analysis of model residuals	296
19.5	COMPARISON OF COX PH REGRESSION MODELS	297
20	RELIABILITY ANALYSIS	305
21	THE WIZARD	311

22 OTHER NOTES	312
22.1 FILES FORMAT	312
22.2 SETTINGS	313



1 SYSTEM REQUIREMENTS

To use PQStat, your computer must meet the following minimum requirements:

- Processor: Intel Pentium II (500 MHz or better)
- 256 MB RAM or greater
- SVGA (800 x 600/16-bit colour or better)
- 200 MB of disc space
- The alternate install CD only requires you to have: CD-ROM
- Other requirements: a keyboard, a mouse
- Supported Operating Systems: Windows 2000/XP/Vista/7/8

2 HOW TO INSTALL

To start the installation process, run the application installer - PQStat-setup_x86-FULL (for 64-bit version: PQStat-setup_x64-FULL.exe).

When you do this, a setup dialog box will appear. Press "Next" to continue with the installation setup. The installation of the application requires you to accept the End User License Agreement. If you accept the terms of the license, select: "I accept the terms of the license" and press "Next" to continue. Otherwise, select "I do not accept the terms of the licence" and press "Cancel" to exit the installation.

The following box enables you to change the default install[®]ation directory and to check if you have sufficient disc space. It is recommended that the default location of instalation is accepted.

If you press "Next", there is a possibility to choose either a full installation of the application or a version not including exemplary data sets. The data sets are used in the User Guide.

Next, the dialog box informs you and gives you the possibility to change the shortcut name, which will be created in Windows Menu Start.

Pressing "Next", you can create a Desktop Shortcut or add a shortcut to the Quick Lunch toolbar. Press "Next" to continue.

The following step is the last one before the installation process starts copying files to your system. This dialog box will show you the summary of installation options chosen so far. To start the installation process, press "Install".

3 WORKING WITH DOCUMENTS


Documents management in this application is based on projects. Each project is a separated file.

A project is an object of the similar meaning to a worksheet, which consists of 3 basic elements:


1. Datasheets (including map sheets and matrixes) - the number of sheets in a given project is limited to 255,
2. Results sheets (reports) - the number of reports in a given datasheet is limited to 1024,
3. Project manager - it enables you to change the name of datasheets and results, add your own descriptions and notes, and export.

It is possible to work on 255 opened projects at the same time. The first one, altogether with an empty sheet, is created automatically (right after the application is launched, and if the appropriate option in the [application settings](#) is selected).


Another projects can be created by:

- File menu → New project (Ctrl+N),
-  button on the toolbar .

Created projects (files with [pqs](#), [pqx](#) extension) can be opened by:

- File→Open project (Ctrl+O),
-  button on the toolbar,
- File→Open recent,
- File→Open examples - it applies to the examples attached to the application,
- drag the project file into the application window,
- by double-clicking the project file.

The project can be saved by:

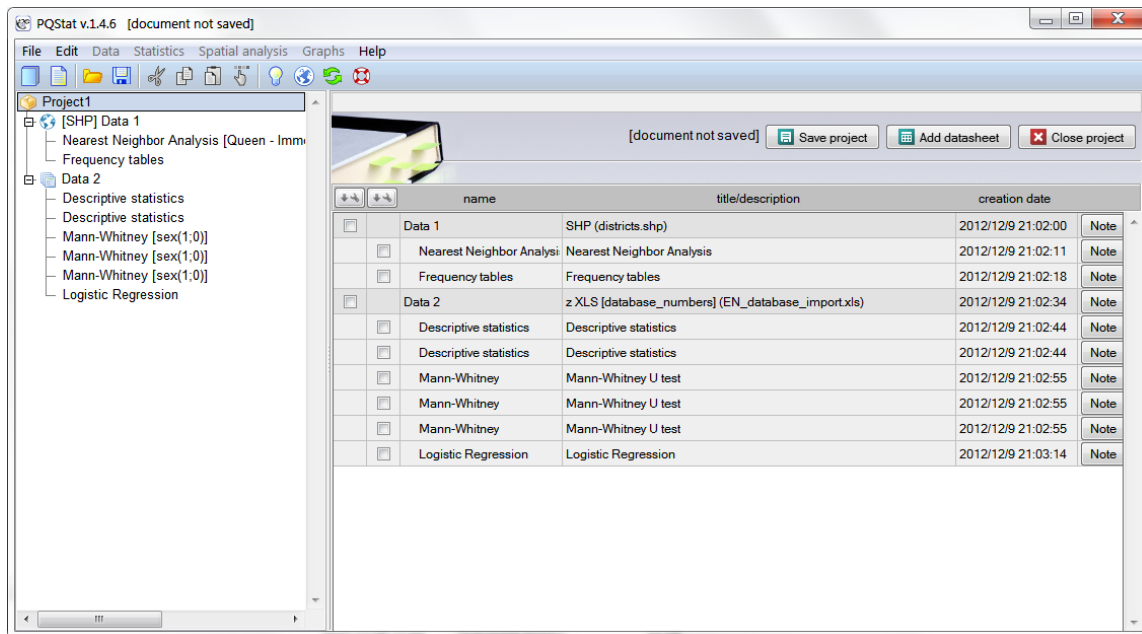
- File menu→Save (Ctrl+S),
- File→Save as...,
- Save button in the Project Manager, -  button on the toolbar.

Saving the project causes that all project elements are saved in a file with [pqs](#) or [pqx](#) extension.

The project can be closed by:

- File menu→Close project,
- Close project button in the [Project Manager](#).

To navigate the project easily, you can use a Project Manager that is opened when you select appropriate project. In this window, you can both save and delete projects. You are also able to delete datasheets and reports or to add descriptions and notes. Project Name is also the name of the project file (pqs / pqx).




3.1 HOW TO WORK WITH DATASHEETS

The most important element in each project is a datasheet. Each open project must contain at least one datasheet.


3.1.1 HOW TO ADD, TO DELETE AND TO EXPORT DATASHEETS

The first empty datasheet will be opened automatically altogether with a new project.

Another datasheets can be added to the project by:


- File menu → Add datasheet (Ctrl+D),
-  button on the toolbar,
- Add datasheet to the [Project Manager](#).

You can delete a datasheet by:

- context menu Delete sheet (Shift+Del) on the name of a datasheet in a [Navigation Tree](#),
-  button → Delete in the [Project Manager](#), for selected sheet/sheets.

However, you should remember: if there are any reports or map added to a datasheet and you delete datasheet, all reports/map attached to it will be deleted too.

Datasheets can be described in the [Project Manager](#) by adding a name, title or a note.

All datasheets created in PQStat can be exported to [csv \(txt\)](#), dbf and xls format. You can do this by clicking  button → Eksport to.. in the [Project Manager](#), for selected sheet/sheets.

3.1.2 HOW TO INSERT DATA INTO A SHEET

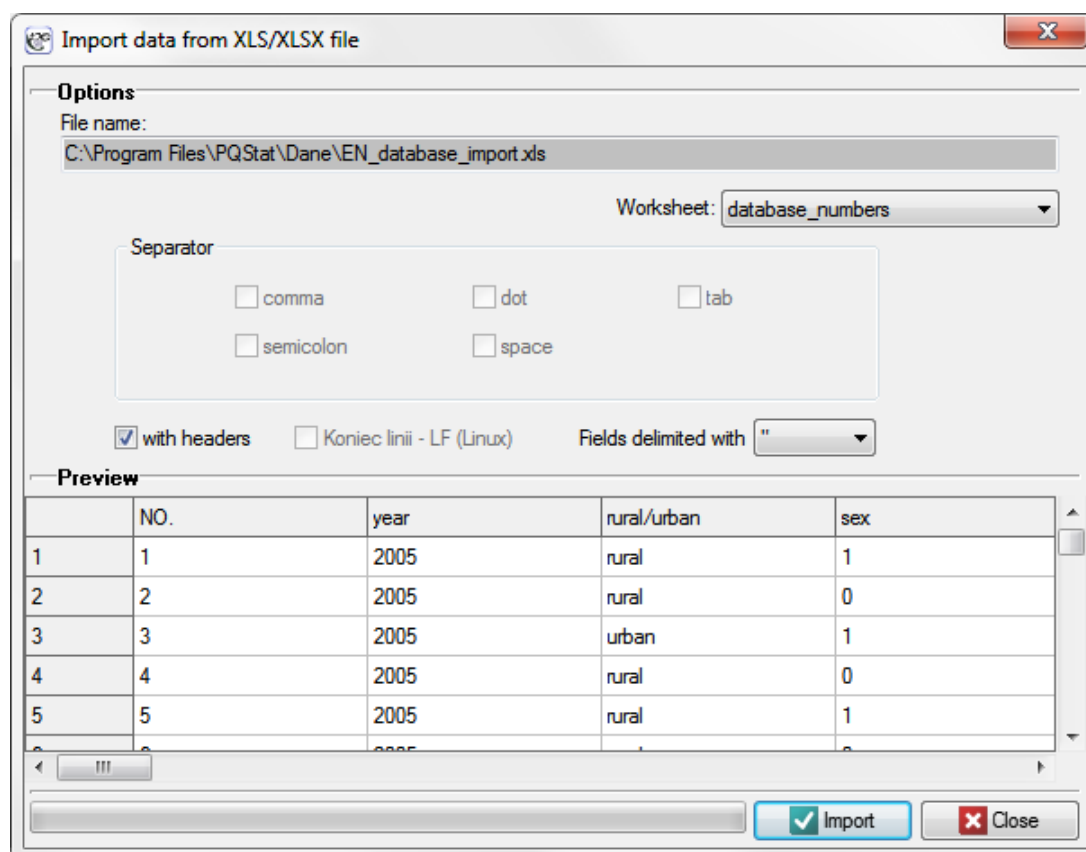
Creating a datasheet, it is empty. You can insert some data, copy prearranged collection of data from any datasheet or import data. The amount of data, which one datasheet is able to take in is limited to 4 millions of rows and 1 thousand of columns. No more than 40 characters can be put in each cell.

Data import

You can easily import data from:

- *.xls/*.xlsx,
- *.txt/*.csv files with encoding of UTF8, Windows-1250,
- *.shp (SHP/SHX/DBF ESRI Shapefile),
- *.dbf (dBase III, dBase IV, dBase VII),
- *.dbf (FoxPro).

To perform an import operation you should click Import from... menu.



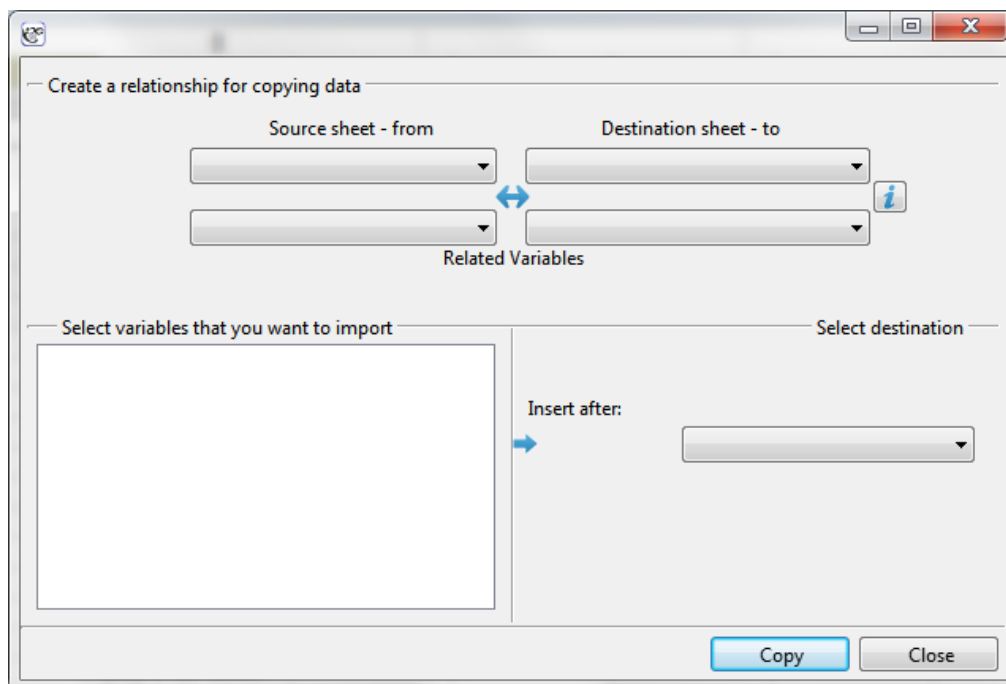
In the import window, there is a possibility to preview data importing and prior verification of import results, depending on the way of data interpretation. To avoid misinterpretation of national characters, you should pay special attention on the correctness of screened characters in a preview window. If the files are huge, the preview window displays only the beginning of the data from the given file.

Note

In applications like Microsoft Office Excell 2000-2007, the default character encoding is Windows-1250. Data importing from Microsoft Excel documents is with reference to cells values only. There is no possibility to import any formatting and formulas.

Copying data with relation

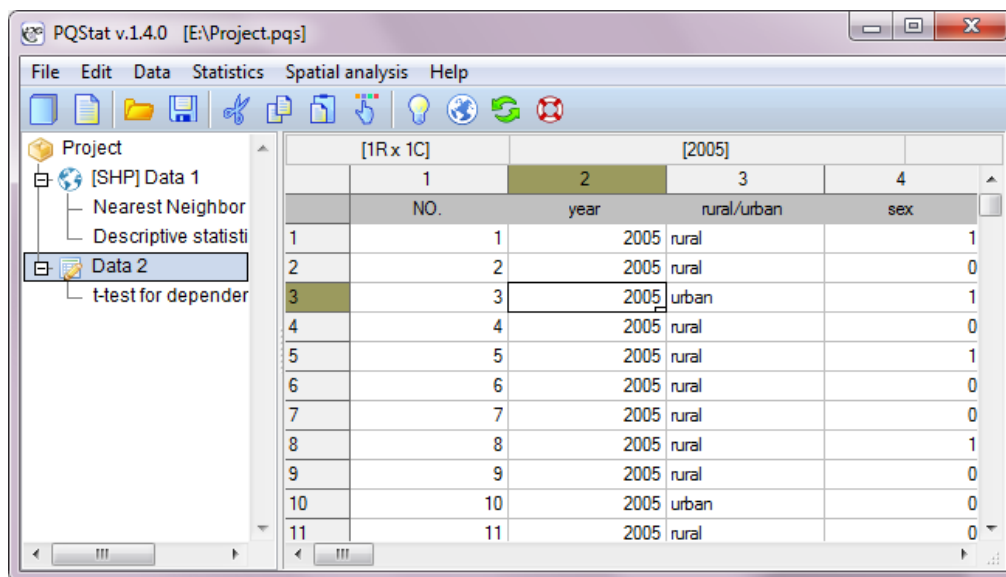
Data from one datasheet can be copied to another selected datasheet on the basis of relation. That kind of copying is done by selecting from the menu Data→Copying with relation...



In order to build a relationship one ought to select the datasheet from which the copying is to be done and the datasheet into which the copied data will be transferred. Both datasheets ought to have the same key, i.e. the variable the values of which identify each row in the datasheet. The key for the source datasheet must be unique. The principle of the design is a one-to-many relationship, i.e. one row from the source datasheet can be related to many rows from the destination datasheet. The keys of both datasheets ought to be selected as Related variables. Having set the relationship as described above, we select the variables to be copied and to the column after which the copied variables are to be placed.

3.1.3 DATASHEET WINDOW

Rows and columns of a datasheet are marked with successive natural numbers. You can give your own header to each column in a place where grey colour occurs. There is a Message bar at the top of each datasheet. The message bar displays all current information for you. The left side of the bar gives you all information about the dimension of the selected area [like the number of rows, columns], the centre part of the bar displays the value occurred in the selected cell and the right side of the bar gives you information mainly about a statistical analysis which is in progress at that moment.



3.1.4 CELLS FORMAT

Each datasheet cell (including the column heading) can contain a maximum of 40 signs. Also allowed are texts containing national characters. The introduced values can be formatted as:

- **default** – in the case of the default format the program automatically recognizes the content of a cell with regard to numerical and text data;
- **text** – in the case of the text format the data are interpreted as text (alignment to the left edge of the cell);
- **data** – in the case of the date format the data are interpreted as subsequent values of a date, thus value 1 means 1899.12.31, value 2 means 1900.01.01, and so on. Depending on the selected date format one can also introduce text data in a selected format:

2010.12.31
 31.12.2010
 12.31.2010
 2010/12/31
 31/12/2010
 12/31/2010
 2010-12-31
 31-12-2010
 12-31-2010

- **time** – in the case of the time format the data are interpreted as subsequent values of time, and the decimal part of a number means the number of milliseconds from midnight divided by the total number of milliseconds in a day (86,400,000), thus value 0.000694444 means 00:01:00, value 0.041666667 means 01:00:00, and value 0.999988426 means 23:59:59. Depending on the selected time format one can also enter text data in a selected format:

18:31:58
 18:31
 12/31/2010 18:31
 12/31/2010 18:31:58

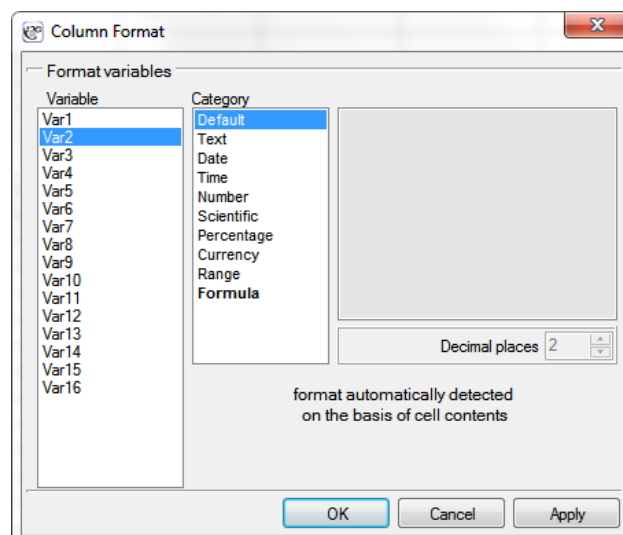
- **numerical** – real numbers in this format are in the form of a decimal, and the sign dividing the whole number from the decimal number is a comma or a dot (depending on the settings selected in the window hyperlinkSettingsSettings in the field Decimal separator), it is possible to set the number of decimals and the thousands separator;
- **scientific** – i.e. when $M \cdot 10^E$ is used, where the basis is the M mantissa, and the E - index of the power is an integer; as in the numerical format it is possible to set the number of decimals;
- **percentage** – they change the number into a percentage by multiplying by 100 and displaying it with the % symbol; as in the case of the numerical format it is possible to set the number of the decimals;
- **currency** – used for money values; allows to add the symbol of a currency; as in the case of the numerical format it is possible to set the number of the decimals;
- **range** – marked with the use of the upper and lower boundary; as in the case of the numerical format it is possible to set the number of the decimals;
- **formula** – values calculated according to the formula ascribed to the column; the values are automatically recalculated when any of the entry data is changed.

When a new sheet is opened, there is a standard default format for each cell. In a default format the sheet supports cell content automatically.

A whole header row is set permanently of the text format. You can set defined formats for the rest of the sheet. Only a whole column can be formatted (except for its header), not a single cell.

To set a column format you should select:


- Format in a context menu of the number displayed above a column header,
- Edit→Column format, when an active cell identifies the proper column.



You can define the **width of a column** by using a mouse arrow. In order to do this, you should move the line which divides two neighbouring columns to narrow or widen the column on the left side of above mentioned line.

Additionally, you can set different colour of the background in each cell of a sheet (when you select the

area you want to change). To do this, use:




-  button on the toolbar,
- Cell colour command on the cell's context menu.

3.1.5 DATA EDITING

You can **select the consistent area of a sheet** using a mouse or a keyboard (Keyboard arrows + Shift). While selecting an area, its size is displayed currently on the Message box (the number of rows and columns). You can easily select the whole sheet by clicking the top left corner of the sheet or selecting from the menu Edit→Select all (Ctrl+A). If you want to select the whole columns or rows, just click their headers.

Cell Copying and moving is performed with Copy, Cut and Paste.

The above commands can be found in several places like:

- Edit menu,
- Context menu of each cell or cells,
-    buttons on the toolbar,
- Context menu of the columns and rows,
- Shortcut keys: Copy (Ctrl+C), Cut (Ctrl+X), and Paste (Ctrl+V).

To **delete data from cells** select Edit→Delete (Del)

If you want to **undo recent operations** select Edit→Undo (Ctrl+Z). There are 10 recent operations automatically saved in a Program memory. Each operation refers to maximum 5000 cells. These settings may be changed in a [Settings window](#). However, note that the higher the values used in a operation, the more computer memory is used by the program.

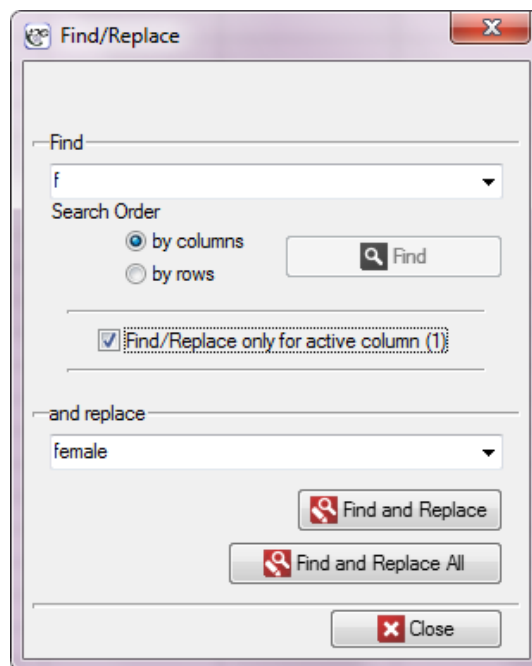
How to insert and delete rows and columns

You can insert empty columns or rows above or on the left side of already existing ones. It will move the old ones down or to the right side. To insert row/rows, you should select the one/ones above which you want to insert new ones. Then, you should choose Insert row in a context menu of the number of selected row. Exactly the same way you can insert new columns.

Rows and columns can be both inserted and deleted. You can delete them by selecting Delete row/Delete column on the context menu of the number of a row or a column.

How to find/replace a cell value

To find or replace cell value contents with another value, you should use a Search/Replace window, which you can find in Edit menu→Find/Replace (Ctrl+F). To search, use upper half of the window, to change a cell content, use lower half of the window.

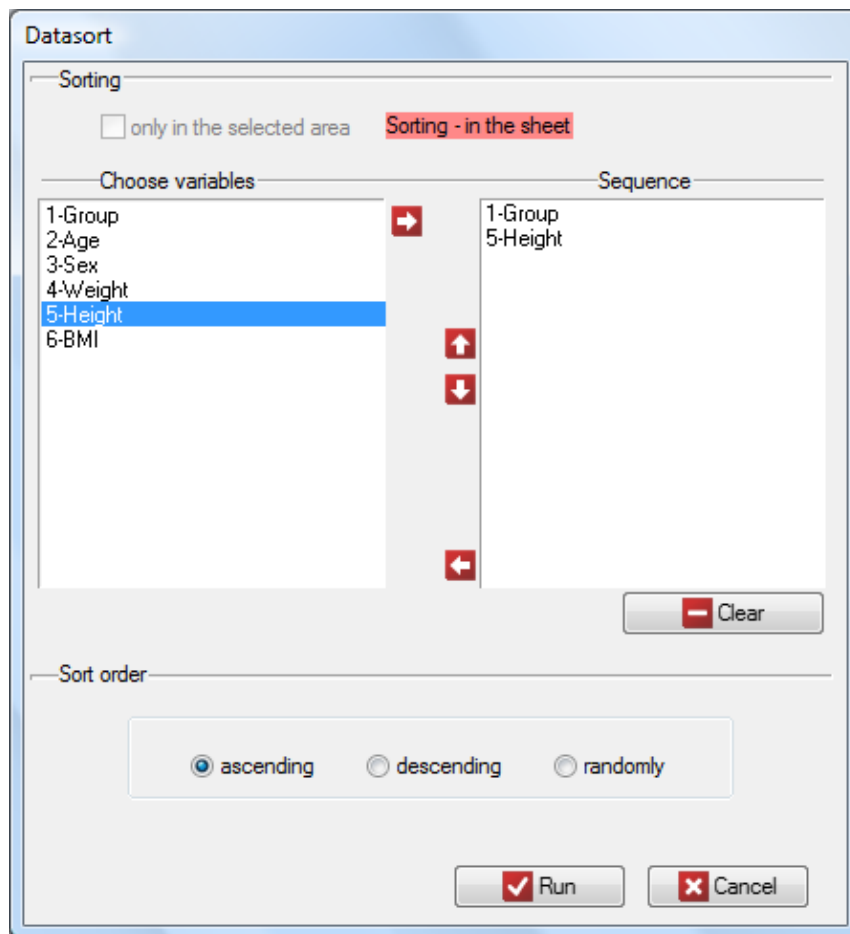


To find specific data, you should write the right characters in the upper half of the window, then select the sequence of searching and click Find.

To find and to replace the whole cell content with another value, you should fill in an upper half as well as a lower half of the window. An upper half should be filled in exactly the same way as you do with data searching. In the lower half of the window you should insert data which are supposed to replace the already found one. Then you should click Find and Replace or Find and Replace All (if you want to replace all the found data which occurred). Both searching and replacing data accompanies a direct preview of a current action on the sheet.

3.1.6 HOW TO SORT DATA

The options of sorting data will be found after choosing Sort... from Data menu or Sort... option in a context menu of the number displayed above a column header. Usually the whole datasheet is sorted (this is a default setting), but if you first select the part of the data, then in the sorting window you will have an opportunity to reduce the area just to this selected part of the data.



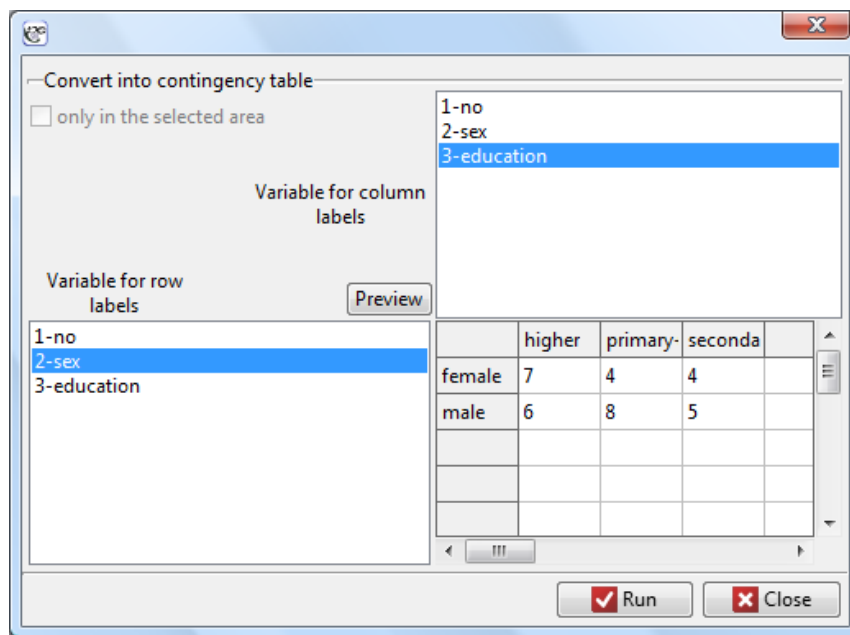
In the window of sorting, you can move (using indicators) from Choose variables box to Sequence box these variables, according to which you want to sort the data. Then you should choose Sort order and confirm your choice by clicking Run.

You can choose maximum 3 columns as a criteria of sorting. If you sort data using more than one criterion, then sorting is performed according to column (variables) sequences, placed in a Sequence box.

3.1.7 HOW TO CONVERT RAW DATA INTO CONTINGENCY TABLE

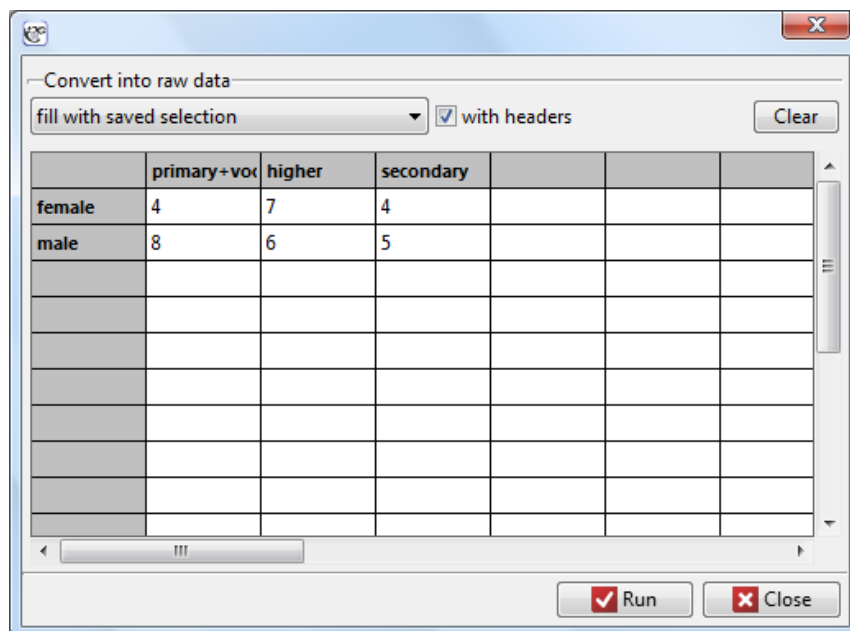
You can start the operation of converting [raw data](#) into a [contingency table](#) by selecting Create table... from Data menu. Usually, there is the whole data sheet available for this operation (default). However, if you start the conversion from selecting a piece of data, you will be able to reduce the area available only to the selection.

A contingency table can be designed by selecting the variables forming row and column labels. If a preview of the table does look like the expected one, you confirm the choice by selecting Run. The returned result will be placed in a new datasheet.



3.1.8 HOW TO CONVERT CONTINGENCY TABLE INTO RAW DATA

You can start the operation of converting a [contingency table](#) into [raw data](#) by selecting Create raw data... from Data menu. In the window of data transformation, we enter appropriate numbers and headers of rows and columns. You confirm the choice by selecting Run. The returned result will be placed in a new datasheet.

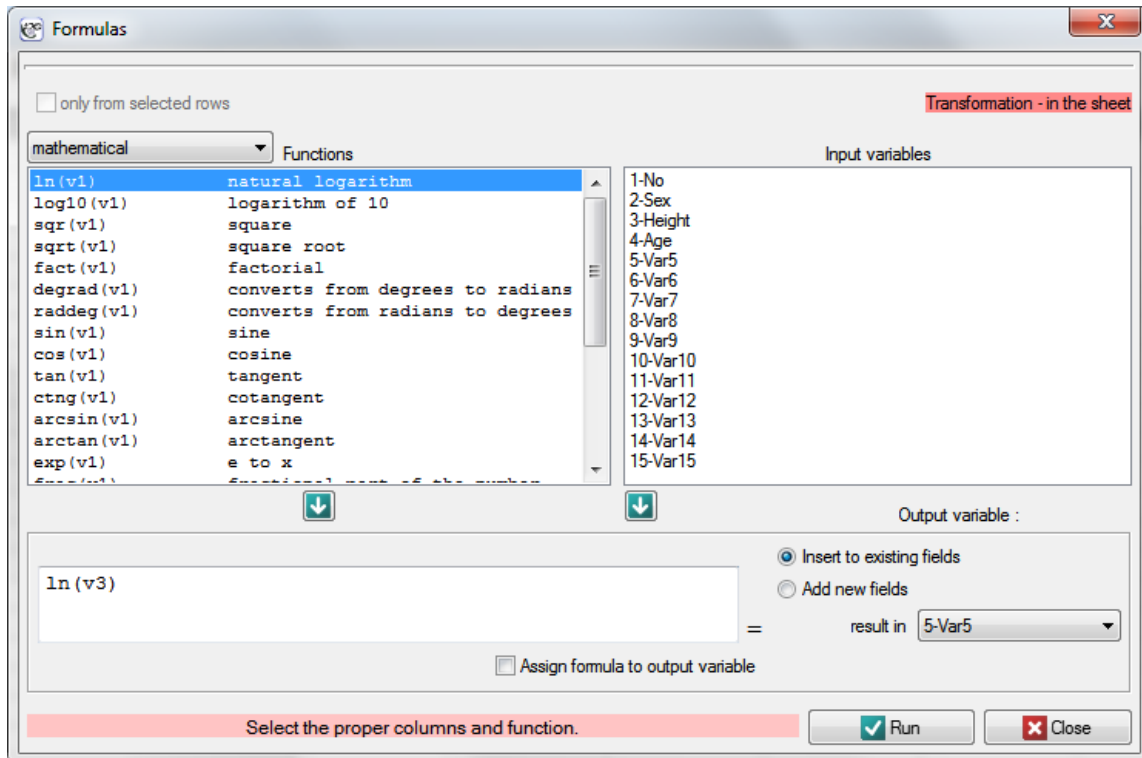


If we convert a table which is placed in a datasheet, we have to select it (with or without header) before the conversion of the table into raw data. Then, in the conversion window, the table will be placed automatically. It is also possible to use other labeled tables as a [saved selection](#).

3.1.9 FORMULAS

Defining the formula is a way of calculating data so as to obtain new values for the variables.

The window in which we define formulas is accessed by selecting Data→Formulas...



Formulas ascribed to a given variable of the datasheet as the format of that variable are remembered together with the datasheet. Their result is automatically recalculated when any of the entry data are changed. The formula can be ascribed in the Formulas... window or by selecting Column format (Ctrl+F10).

Building formulas

We write formulas in the edition field.

- We enter the variables to which the formulas refer by giving their numbers, e.g. $v1+v2$.
- Text values are entered with the use of an apostrophe, e.g. 'house'.
- We enter functions by double clicking on the name of the selected function. The name then appears in the edition field of the formula. Alternatively, we can enter the name directly in the edition field. In such a case the capitalization of the letters in the name of the function does not matter. The function arguments are given in brackets, with the use of the syntax given in the description of the function,

Formula results

The results of the formulas will be displayed in the selected column.

If among the arguments of the function there will be values which the function cannot interpret, the program will display a message asking whether the uninterpreted data ought to be omitted. A confirmation will cause a recalculation of the formula without the uninterpreted data. If a negative answer is given, the error value NA will be returned. For example, for values in columns $v1$, $v2$, and $v3$, respectively: 1, 2, 'ada', the sum function $\text{sum}(v1; v2; v3)$ will return the result 3 if we skip the uninterpreted value 'ada' or will return NA if we do not skip that value in the calculations.

An empty value (missing data) will only be returned when all the arguments used in the formula are

empty.

The number of rows taking part in the formula can be limited by selecting an appropriate range of rows in the datasheet and by selecting the option only from selected rows in the formula window.

Operators

- + addition,
- − subtraction,
- * multiplication,
- / division,
- % modulo division (as a result the remainder of division of one number by another),
- > greater,
- < lower,
- = equal.

Mathematical functions

Mathematical functions require numeric arguments.

- ln(v1)** - returns a natural logarithm of the given number,
- log10(v1)** - returns a logarithm to the base 10 of the given number,
- logn(v1)** - returns a logarithm to the base n of the given number,
- sqr(v1)** - returns a value of the given number raised to the 2nd power,
- sqrt(v1)** - returns a value of the square root of the given number,
- fact(v1)** - returns a value of factorial of the given number,
- degrad(v1)** - returns the angle in radians (argument are degrees),
- raddeg(v1)** - returns the angle in degrees (argument are radians),
- sin(v1)** - returns sinus of the given angle, (argument are radians),
- cos(v1)** - returns cosinus of the given angle, (argument are radians),
- tan(v1)** - returns tangens of the given angle, (argument are radians),
- ctng(v1)** - returns cotangens of the given angle, (argument are radians),
- arcsin(v1)** - returns arcus sinus of the given angle, (argument are radians),
- arctan(v1)** - returns arcus tangens of the given angle, (argument are radians),
- exp(v1)** - returns e raised to the power of the given number,
- frac(v1)** - returns the fractional part of the given number,
- int(v1)** - returns the integer part of the given number,
- abs(v1)** - returns absolute value of the given number,
- odd(v1)** - returns 1 if the given nummber is even or 0 if the given number is odd,
- sum(v1;...)** - returns the result of an addition of the given numbers,
- multip(v1;...)** - returns the result of a multiplication of the given numbers,
- power(v1;n)** - returns a value of the given number raised to the n -th power,
- norme(v1;...)** - returns the Euclidean vector norm,
- round(v1;n)** - returns a number rounded to n decimal places.

Statistical functions

Funkcje statystyczne wymagają argumentów liczbowych.

- stand(v1)** - returns a standardised score of the given numbers,
- max(v1,...)** - returns the highest value out of the given numbers,
- min(v1,...)** - returns the lowest value out of the given numbers,
- mean(v1,...)** - returns the arithmetical mean value of the given numbers,
- meanh(v1,...)** - returns the harmonic mean value of the given numbers,
- meang(v1,...)** - returns the geometric mean value of the given numbers,

median(v1,...) - returns the median value of the given numbers,
q1(v1,...) - returns the lower quartile of the given numbers,
q3(v1,...) - returns the upper quartile of the given numbers,
cv(v1,...) - returns the coefficient of variability value of the given numbers,
range(v1,...) - returns the range value of the given numbers,
iqrange(v1,...) - returns the interquartile range value of the given numbers,
variance(v1,...) - returns the variance value of the given numbers,
sd(v1,...) - returns the standard deviation value of the given numbers.

Text functions

Text functions work on any string of characters.

upperc(v1) – converts the characters from the string into capitalized characters,
lowerc(v1) – converts the characters from the string into characters written with small letters,
clean(v1) – removes the unprintable signs,
trim(v1) – removes initial and final spaces,
length(v1) – returns the length of the string of characters,
search('abc';v1) – returns to the beginning of the search string
concat(v1;...) – joins texts,
compare(v1;...) – compares texts,
copy(v1;i;n) – returns a part of the text, starting from the *i*th character, where *n* is the number of the returned characters,
count(v1;...) – returns the number of cells which are not empty,
counte(v1;...) – returns the number of empty cells,
countn(v1;...) – returns the number of cells which contain numbers.

Date and time functions

The date and time functions should be performed on data formatted as date or as time (see chapter 3.1.4). If that is not the case, the program tries to recognize the format automatically. When that is not possible it returns the NA value.

year(v1;) – returns the year ascribed to the date,
month(v1;) - returns the month ascribed to the date,
day(v1;) - returns the day ascribed to the date,
hour(v1;) - returns the hours ascribed to the time,
minute(v1;) - returns the minutes ascribed to the time,
second(v1;) - returns the seconds ascribed to the time,
yeardiff(v1;v2) - returns the difference in years between two dates,
monthdiff(v1;v2) - returns the difference in months between two dates,
weekdiff(v1;v2) - returns the difference in weeks between two dates,
daydiff(v1;v2) - returns the difference in days between two dates,
hourdiff(v1;v2) - returns the difference in hours between two times,
minutediff(v1;v2) - returns the difference in minutes between two times,
seconddiff(v1;v2) - returns the difference in seconds between two times,
compdate(v1;v2) - compares two dates and returns the number 1 when $v1 > v2$, 0 if $v1 = v2$, -1 if $v1 < v2$.

Logical functions

if(question;'yes answer';'no answer') – the question has the form of a statement which can be true or false. The function returns one value if the statement is true and another value if it is false,
and – conjunction operator – returns the truth (1) when all the conditions it connects are true;

otherwise, it returns falsity (0),

or – alternative operator – returns the truth (1) when at least one of the conditions it connects is true; otherwise, it returns falsity (0),

xor – either/or operator – returns the truth (1) when one of the conditions it connects is true, otherwise, it returns falsity (0),

not – negation operator – used in a conditional sentences if.

3.1.10 HOW TO GENERATE DATA

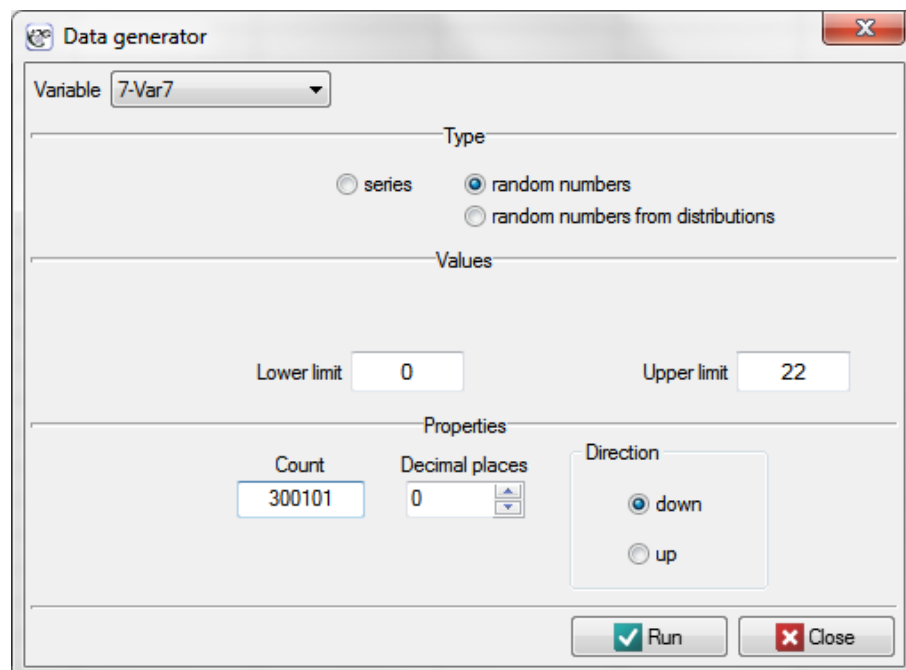
There are 2 methods of data generation:

1. The first method uses a pull technique. All the data are pulled from the selected cells into the neighbouring ones using a mouse arrow. This method enables you to generate exactly the same values (number or text ones) in the neighbouring columns or rows.

To start data generation, select a cell with the proper content, then click on the right down corner using a mouse arrow illustrative + sign and not letting it go just pull through all the cells you want to fill. Pulling one cell can be done in any direction (up, down, right, left). It is also possible to pull various values which are put in a one column (left or right) or in a one row (up or down).

2. The other method enables you to generate numerical data in columns as: a data sequence, random values or random values of the proper data distribution.

To generate numerical data you should select a cell, where you want to start filling the datasheet and open data generation window in Data menu→Generate...



We indicate a variable, in which the generated data will be placed.

In the middle part of the window, depending on the way of data generation settings chosen above, set:

- To generate data series:
 - Start value - the first value which needs to be generated,

- Increment - a value which is supposed to be the difference between the following generated data.
- To generate random numbers:
 - Lower limit - beginning of the interval, from which the values will be randomised,
 - Upper limit - end of the interval, from which the values will be randomized.
- To generate random values from the distribution, you should choose the sort of distribution (Normal distribution, Chi-square distribution) and then write its parameters.

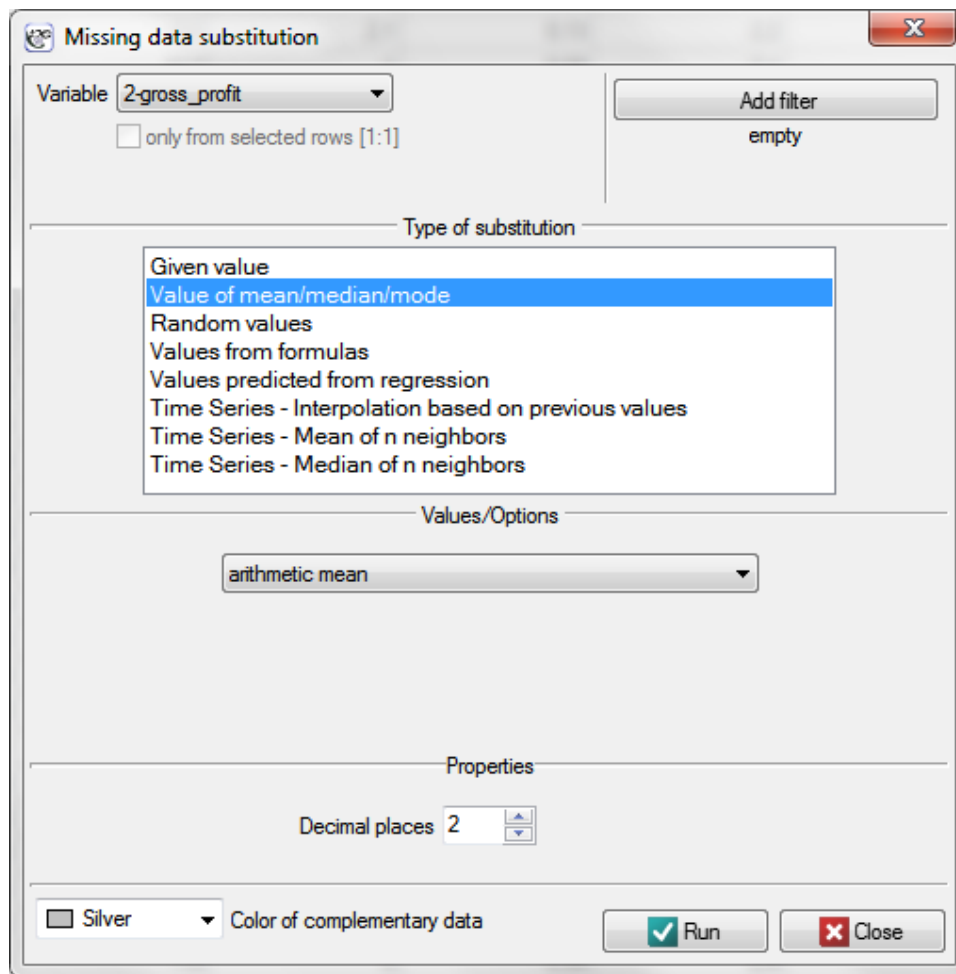
The amount of generated data depends on the value you put in the Count field, but the precision depends on settings of the Decimal places field. Data will be put up or put down starting with an active cell - it depends on a selected option. At the end, confirm your choice by clicking Run.

3.1.11 MISSING DATA

In studies we very often see missing data. That is especially to be expected in the case of survey data. There are situations in which the missing data gives valuable information. For example, the number of missing data in answer to a question concerning preferences with regard to political parties informs us about the number of undecided citizens who do not favor (or do not admit they do) particular political groups. Small amounts of missing data do not constitute a problem in statistical analyses. Large amounts, however, can undermine the reliability of the conducted research. It is worth taking care that there are as few such lacks as possible, from the start. Obviously, it would be preferable to gain access to the real value and enter it in place of the missing data but that is not always possible.

The manner in which the missing data are treated depends, primarily, on their character. In this program a number of ways have been implemented for imputing the missing data for particular variables.

The window with the settings for the replacing missing data option is accessed from the menu Data→Missing data...



1. Filling in with one value

Selecting one of the options below will cause the replacement of all the missing data in the selected column it with the same value.

- given by the user,
- the arithmetic mean calculated from the data,
- the geometric mean calculated from the data,
- the harmonic mean calculated from the data,
- the median,
- the mode (unless it is multiple).

2. Filled with many values

The selection of one of the options below will cause the replacement of the missing data in the selected column with many (usually different) values. The values can be predicted on the basis of the column for which the missing data are being replaced or on the basis of the values of other columns (variables). The missing data can be replaced with the following types of values:

- random values from the dataset,
- random values from the normal distribution defined on the basis of the mean and the standard deviation from the existing data,

- random values from a range given by the user,
- calculated from the user's functions, which allows the use of data from other variables so as to be able to predict the missing value in the selected column,
- calculated from the regression model, which allows to predict the values of the missing data on the basis of a multiple regression model (the manner in which multiple regression operates was described in chapter ?? [Multiple linear regression](#)),
- interpolation on the basis of the neighboring values – it applies to time series – so the user must point to the time variable which gives information about the data order; the interpolation consists in the determination of the value for the missing data in such a manner that they are placed, graphically, on a straight line joining the values of the data neighboring the missing data,
- the mean from the n of the neighbors – it applies to time series – so the user must point to the time variable which informs about the order of data; the interpolation consists in determining a mean from the value for n antecedent neighbors and n neighbors directly following the missing data,
- the median from n neighbors – it applies to time series – so the user must point to the time variable which informs about the order of data; the interpolation consists in determining a median from the value for n antecedent neighbors and n neighbors directly following the missing data.

Note!

In order to be able to distinguish the imputed data from the real data, the replaced data are marked with a selected color.

EXAMPLE 3.1. (file: missingData - publisher.pqs)

The analysis of the file *wydawca.pqs* not containing missing data was discussed in the chapter [Multiple linear regression](#). This time we will discuss a datasheet in which, in the column containing the gross profit from a sale of books, there are missing data. In the case of those missing data we know the real values (datasheet: "REAL VALUES") so we can refer the values generated in the program in the place of the missing data to the real values and compare the results obtained with the use of various techniques. In the example we will use 2 methods of replacing missing data: replacing them with the value of the median and replacing them with a value determined on the basis of a regression model. The remaining possibilities can be studied independently.

Replacing the missing data with the value of the median is done with the use of the first datasheet called "Insert the median". In the Missing data window we set a variable filled in as the gross profit and in this way select the value of the median as a method of replacement. Consequently, the missing data will be replaced with the value USD 46,850.

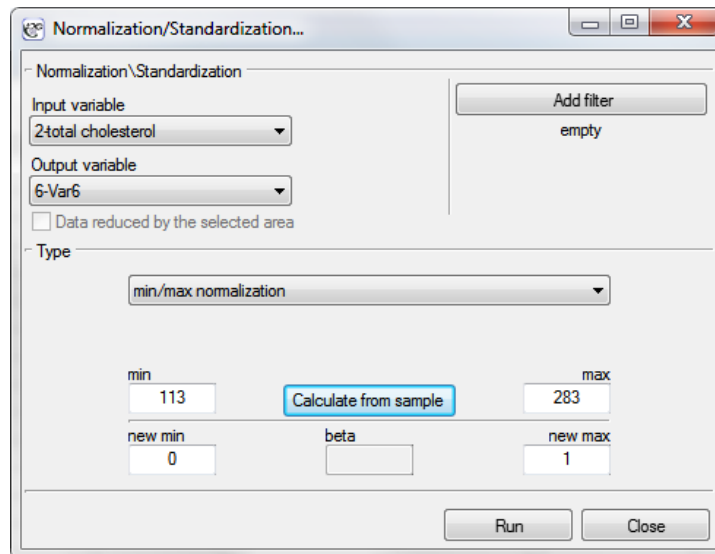
We suspect that the profits are greater when famous authors' books (coded as 1) are sold and smaller when they arise from the sale of less known authors' books (coded as 0). We will, then, calculate the median of the gross profit separately for the famous authors' books and for the less known authors' books. The imputation is made on the datasheet called "Insert two medians". We set, twice, a filter for the variable defining the popularity of an author (variable 7), giving it, respectively, values 1 and 0. The obtained median of the gross profit in the group of the popular authors' books is about USD 51,000 and in the group of the less popular authors' books it is about USD 34,000.

The missing data can also be replaced with the use of the regression model. We choose the data sheet "Insert from regression" and once more select, in the Missing data window, a variable concerning the gross profit as the variable which ought to be filled in, and select the Values predicted from regression

as a replacement method. This time there will be more variables allowing us to predict the value of the gross profit. They will be: production costs (variable no.3), advertising costs (variable no.4), and author's popularity (variable no. 7). The results now seem to be less distant from the real values. However, there is no result for position no. 35, because there was no information about the production costs of that book, that is the factor on which we wanted to base our prediction.

3.1.12 NORMALIZATION/STANDARDIZATION

The normalization/standardization window is accessed via Data→Normalization/Standardization...



The normalization of data is scaling them to a range, e.g. to a range of [-1, 1] or [0,1].

Min-max normalization

The min-max normalization with the use of a linear function scales data to a (new_{min} , new_{max}) range indicated by the user. For that purpose we should know the range which the data can reach. If we do not know the range we can avail ourselves of the greatest and the smallest values in the analyzed set (in such a case we select the calculate from sample option in the Normalization/Standardization window).

$$x' = \frac{x - \min}{\max - \min} \cdot (new_{\max} - new_{\min}) + new_{\min} \quad (1)$$

Logarithmic normalization

Normalization with the use of the logarithmic function (S-shaped) reduces the data to the range of (0,1).

$$x' = \frac{e^x}{1 + e^x} \quad (2)$$

If we want to extend the transformed data in a different range then we ought to enter, in the Normalization/Standardization window, the limits of the new range.

Normalizing function with a coefficient

The normalization reduces the data to the range of (-1,1) with the use of an S-shaped function with the changing α normalization coefficient.

$$x' = \frac{x}{\sqrt{x^2 + \alpha}} \quad (3)$$

When the value of the α coefficient is raised, a graph with a less steep slope is formed.

If we want to extend the transformed data in a different range then we ought to enter, in the Normalization/Standardization window, the limits of the new range.

Standardization

Standardization is the transformation of data as a result of which the mean of a variable is equal to 0 and its standard deviation is equal to 1.

$$x' = \frac{x - \bar{x}}{sd} \quad (4)$$

EXAMPLE 3.2. (file: normalization.pqs)

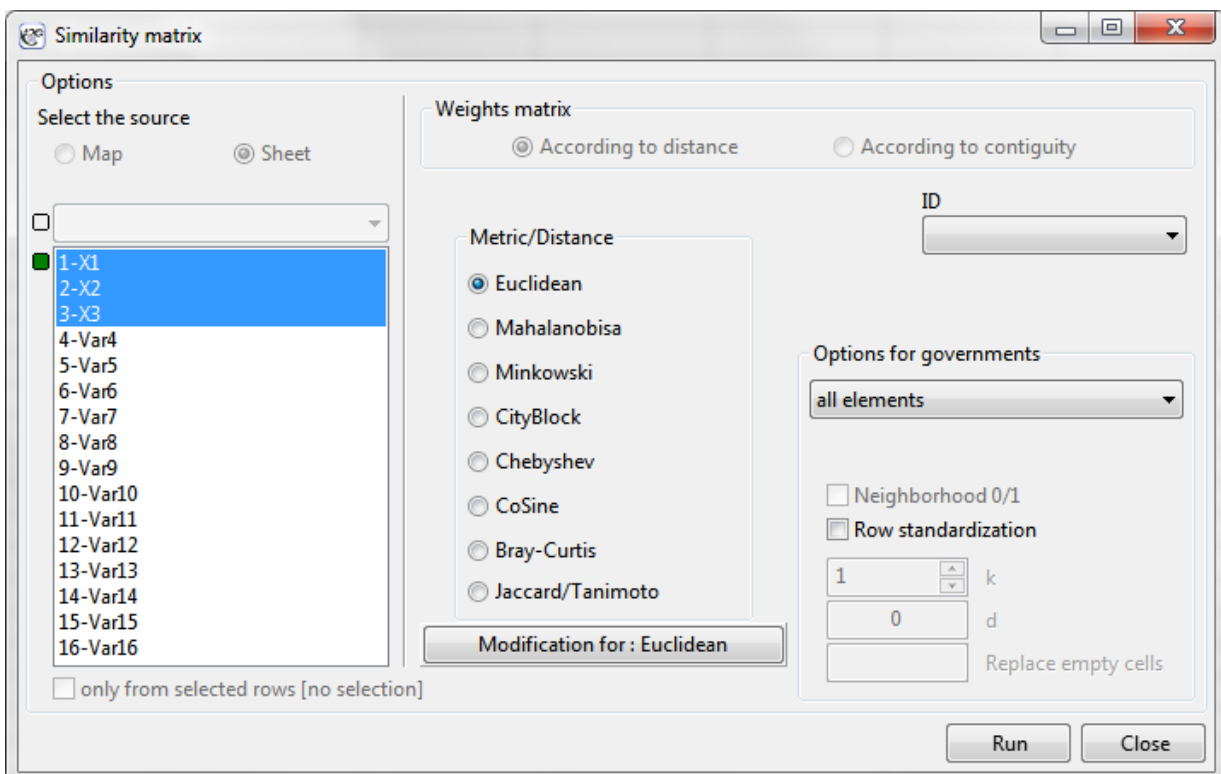
Make the transformations of all the variables included in the file

- using the minimum-maximum normalization to the range [0.10];
- using the logarithmic normalization;
- using the normalization with a coefficient;
- using standardization.

3.1.13 SIMILARITY MATRIX

The mutual relationships among objects can be expressed by their distances or, more generally, by the differences among them. The further from one another the objects are the more they differ, the closer they are, they resemble one another. One can study the distance of the objects with respect to many features, e.g. when the compared objects are cities, we can define their similarity on the basis of, among other things: the length of the road which joins them, population density, GDP, pollution emissions, average property prices, etc. With so many characteristics at the researcher's disposal, he or she must select such a measure of distance as will best represent the real similarity of objects.

The window with the settings for the similarity matrix option is accessed from the menu Dane→Similarity matrix...



Similarity matrix

Options

Select the source

☐ Map ☒ Sheet

☐ [Dropdown]

- ☒ 1-X1
- ☐ 2-X2
- ☐ 3-X3
- ☐ 4-Var4
- ☐ 5-Var5
- ☐ 6-Var6
- ☐ 7-Var7
- ☐ 8-Var8
- ☐ 9-Var9
- ☐ 10-Var10
- ☐ 11-Var11
- ☐ 12-Var12
- ☐ 13-Var13
- ☐ 14-Var14
- ☐ 15-Var15
- ☐ 16-Var16

☐ only from selected rows [no selection]

Weights matrix

☒ According to distance ☐ According to contiguity

Metric/Distance

- ☒ Euclidean
- ☐ Mahalanobisa
- ☐ Minkowski
- ☐ CityBlock
- ☐ Chebyshev
- ☐ CoSine
- ☐ Bray-Curtis
- ☐ Jaccard/Tanimoto

Modification for : Euclidean

ID: [Dropdown]

Options for governments

[Dropdown: all elements]

☐ Neighborhood 0/1

☒ Row standardization

k: [1] d: [0]

☐ Replace empty cells

Run **Close**

The differences/similarities of the objects are expressed with the use of distance, usually in the form of a **metric**. However, not every measure of distance is a metric. For a distance to be called a metric it has to fulfill 4 conditions:

1. the distance between the objects cannot be a negative number: $d(x_1, x_2) \geq 0$,
2. the distance between the objects equals 0 if and only if the objects are identical: $d(x_1, x_2) = 0 \iff x_1 = x_2$,
3. the distance must be symmetrical, i.e. the distance from the object x_1 to x_2 must be the same as from the object x_2 to x_1 : $d(x, y) = d(y, x)$,
4. the distance must fulfill the conditions of the triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

Note!

The metrics ought to be calculated for characteristics with the same range of values. Otherwise, the characteristics with higher ranges would have a greater influence on the obtained similarity result than those with lower ones. For example, when calculating the similarity of people we can base the calculation on such features as weight or age. Then, the weight in kilograms, in the range from 40 to 150 kg, will have a greater influence on the result than age in the range of 18 to 90 years. For the influence of all characteristics on the obtained similarity result to be balanced we ought to [normalize/standardize](#) each of them before commencing the analysis. If we want to decide on the degree of that influence by ourselves, we should enter our own weights, selecting the type of the metric, after the standardization.

Distance/Metric:

Euclidean

When we talk about distance without defining its type we assume that it is the Euclidean distance, the most popular type of distance, constituting a natural element of models of the real world. The Euclidean distance is a metric described by the formula:

$$d(x_1, x_2) = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2}$$

Minkowski

The Minkowski distance is defined for parameters p and r equal to each other. It is then a metric. Such a kind of a metric allows the control of the process of calculating the similarity by giving values p and r in the formula:

$$d(x_1, x_2) = \sqrt[p]{\sum_{k=1}^n |x_{1k} - x_{2k}|^r}$$

When we increase the r parameter, we increase the weight ascribed to the difference between the objects for every characteristic. When we change the p parameter, we increase/decrease the weight ascribed to less/more distant objects. If r and p are equal to 2 the Minkowski distance comes down to the Euclidean distance. If they are equal to 1 – to the city block distance. If the parameters tend to infinity – to the Chebyshev metric.

city block (also called the Manhattan or taxicab metric)

It is the distance which allows the movement only within two perpendicular directions. That kind of distance reminds movement along perpendicular streets (a square street network reminiscent

of the grid layout of most streets on the island of Manhattan). The metric is calculated with the formula:

$$d(x_1, x_2) = \sum_{k=1}^n |x_{1k} - x_{2k}|$$

Chebyshev

The distance between the compared objects is the greatest of the obtained distances for the particular characteristics of those objects.

$$d(x_1, x_2) = \max_k |x_{1k} - x_{2k}|$$

Mahalanobis

The Mahalanobis distance is also called statistical distance. It is weighted by the covariance matrix, which allows the comparison of objects described by mutually correlated features. The use of the Mahalanobis distance has two basic advantages:

- 1) The variables for which greater deviations or value range are observed do not have an increased influence on the result of the Mahalanobis distance (because when we use a covariance matrix we standardize the variables with the use of the variance on the diagonal). As a result, before starting the analysis one does not have to standardize/normalize the variables.
- 2) It takes into account the mutual correlation of the features describing the compared objects (when we use a covariance matrix we use the information about the dependency among the features, which is placed beyond the diagonal of the matrix).

$$d(x_1, x_2) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$$

The measure calculated in that manner fulfills the requirements of being a metric.

Cosine

The cosine distance ought to be calculated on positive data because it is not a metric (it does not fulfill the first condition: $d(x_1, x_2) \geq 0$). If, then, there are characteristics which also have negative values, we should transform them in advance, with the use, for example, of normalization to a range of positive numbers. The advantage of that distance is that (for positive arguments) it is limited to the range of $[0, 1]$. A similarity of two objects is represented by the angle between the two vectors representing the characteristics of those objects.

$$d(x_1, x_2) = 1 - K,$$

where K is the similarity coefficient (the cosine of the angle between two normalized vectors):

$$K = \frac{\sum_{k=1}^n x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^n x_{1k}^2} \sqrt{\sum_{k=1}^n x_{2k}^2}}$$

The objects are similar if the vectors overlap. In such a case, the cosine of the angle (similarity) equals 1, and the distance (difference) equals 0. The objects are different if the vectors are perpendicular. In such a case the cosine of the angle (similarity) equals 0. The distance (difference) equals 1.

Bray–Curtis

The Bray-Curtis distance (the measure of dissimilarity) ought to be calculated on positive data as it is not a metric (it does not fulfill the first condition): $d(x_1, x_2) \geq 0$. If, then, there are characteristics which also have negative values, we should transform them in advance, with the use, for example, of normalization to a range of positive numbers. The advantage of that distance is the fact that (for positive arguments) it is limited to the $[0, 1]$ range, where 0 means that the compared objects are similar, and 1 – that they are dissimilar.

$$d(x_1, x_2) = \frac{\sum_{k=1}^n |x_{1k} - x_{2k}|}{\sum_{k=1}^n (x_{1k} + x_{2k})} \quad (5)$$

Calculating the measure of similarity BC we subtract the Bray-Curtis distance from value 1:

$$BC = 1 - d(x_1, x_2) \quad (6)$$

Jaccard

The Jaccard distance (measure of dissimilarity) is calculated for binary variables (Jaccard, 1901), where 1 means the presence of a given characteristic and 0 means the absence of it.

		objekt 1	
		1	0
object 2	1	a	b
	0	c	d

The Jaccard distance is expressed with the formula:

$$d(x_1, x_2) = 1 - J. \quad (7)$$

where:

$$J = \frac{a}{a+b+c} - \text{Jaccard's similarity coefficient.}$$

Jaccard's similarity coefficient is within the range $[0,1]$ where 1 means the highest and 0 the lowest similarity. The distance (dissimilarity) is interpreted in the opposite manner: 1 means that the compared objects are dissimilar and 0 that they are very similar. The meaning of Jaccard's similarity coefficient can be illustrated very well by the situation of clients choosing products. The fact of the purchase of a given product by a client will be marked with 1 and the fact of not purchasing the product by 0. When calculating Jaccard's coefficient we will compare 2 products so as to learn how many clients buy them together. We are not, of course, interested in the clients who did not buy any of the compared products. What we are interested in is how many people who bought one of the compared products also bought the other one. The sum $a + b + c$ is the number of clients who bought one of the compared products and a is the number of customers who bought both products. The higher the coefficient the more interrelated the purchases (the purchase of one product is accompanied by the purchase of the other one). The opposite is true if we obtain a high Jaccard's dissimilarity coefficient. Such a situation shows that the products compete with each other, i.e. the purchase of one product will exclude the purchase of the other one.

The formula of Jaccard's similarity coefficient can also be presented in the general form:

$$J = \frac{\sum_{k=1}^n x_{1k} x_{2k}}{\sum_{k=1}^n x_{1k}^2 + \sum_{k=1}^n x_{2k}^2 - \sum_{k=1}^n x_{1k} x_{2k}}$$

proposed by Tanimoto (1957). An important feature of the Tanimoto formula is that it can also be calculated for continuous characteristics.

In the case of binary data, Jaccard's and Tanimoto's dissimilarity/similarity formulas are identical and fulfill the conditions of a metric. For continuous variables the Tanimoto formula is not a metric (does not fulfill the conditions of the triangle inequality).

Example – a comparison of species

We compare the genetic similarity of the representatives of three different species, in terms of the number of genes common to all the species. If a gene is present in an organism, we ascribe it value 1. In the opposite case we ascribe it value 0. For the sake of simplicity only 10 genes are subjected to the analysis.

GENS	gen1	gen2	gen3	gen4	gen5	gen6	gen7	gen8	gen9	gen10
representative1	0	1	1	1	1	1	1	0	1	0
representative2	0	0	1	1	1	1	1	0	1	0
representative3	1	0	1	1	0	0	1	0	0	0

The calculated similarity matrix looks as follows:

REPRESENTATIVES	representative1	representative2	representative3
representative1	0	0.857143	0.375
representative2	0.857143	0	0.428571
representative3	0.375	0.428571	0

The most similar representatives are no. 1 and no. 2, and the least similar ones are no. 1 and no. 3. - Jaccard's similarity of representative 1 and representative 2 is 0.857143 which means that the 2 species share a little above the 85- Jaccard's similarity of representative 1 and representative 3 is 0.375 which means that the 2 species share above 37- Jaccard's similarity of representative 1 and representative 3 is 0.428571 which means that the 2 species share above 43

Similarity matrix options are used for selecting the manner in which the elements of the matrix ought to be returned. By default all elements of the matrix are returned in the form in which they have been calculated according to the accepted metric. We can change it by setting:

Matrix elements:

- minimum means that in each row of the matrix only the minimum value and the value on the main diagonal will be displayed;
- maximum means that in each row of the matrix only the maximum value and the value on the main diagonal will be displayed;
- k of the minimum means that in each row of the matrix as many smallest values will be displayed as indicated by the user who gives the k value and the value on the main diagonal;
- k of the maximum means that in each row of the matrix as many greatest values will be displayed as indicated by the user who gives the k value and the value on the main diagonal;
- elements below d means that in each row of the matrix only those elements will be displayed the value of which will be smaller than value d indicated by the user and the value on the main diagonal;

- elements above d means that in each row of the matrix only those elements will be displayed the value of which will be greater than value d indicated by the user and the value on the main diagonal;

Neighborhood 0/1

By choosing the option Neighborhood 0/1 we replace the values inside the matrix with value 1 and the empty places with value 0. In that manner we indicate, for example, if the objects are neighbors (1) or not (0).

Standardization by rows

Standardization by rows means that each element of the matrix is divided by the sum of the row of the matrix. As a result, the obtained values are in the range from 0 to 1.

Replace the empty elements

The option Replace the empty elements allows the entry of the value which is to be placed in the matrix instead of possible empty elements.

The selected identifier of the object allows us to name the rows and columns of the similarity matrix according to the nomenclature stored in the indicated variable.

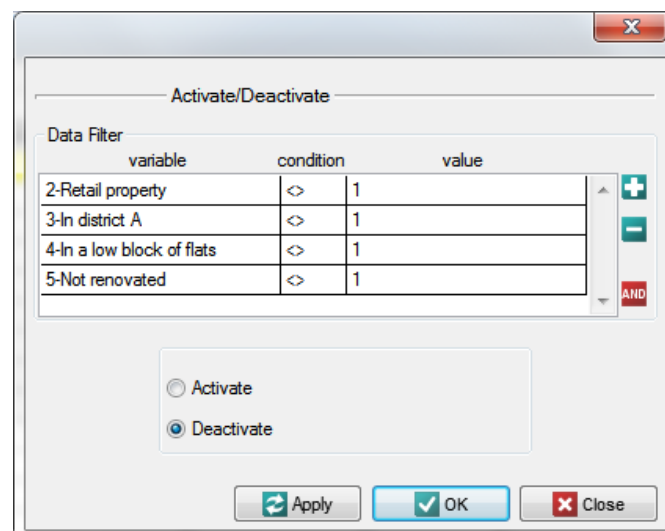
EXAMPLE 3.3. (file: flats similarities.pqs)

In the procedures of property pricing the issue of similarity is very important, for both substantial and legal reasons, For example, it is the main premise for grouping objects and ascribing them to an appropriate segment.

Let us assume that a person who is looking for a flat comes to a real estate agent and defines the obligatory and optional characteristics of the desired property. The characteristics which the flat must have are:

- it is a retail property (the subject of separate ownership),
- it is in district A,
- it is located in a low block of flats (a maximum of 5 floors),
- it is not renovated (average standard or sub-standard).

The data concerning those flats are gathered in a table where 1 means that the property fulfills the search conditions and 0 means that it does not fulfill them.[0.2cm] The flats which do not fulfill the search conditions will be excluded from the analysis by deactivating appropriate rows. We deactivate the rows which do not fulfill any of the conditions, in the menu Edition→Activate/Deactivate (filter)....



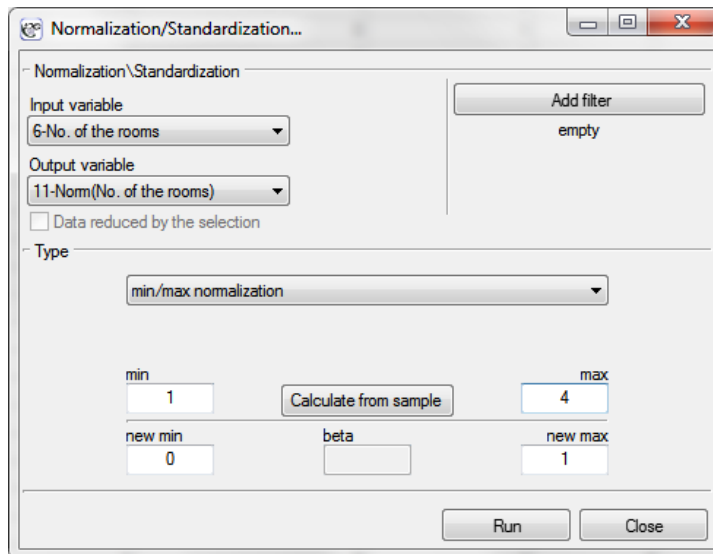
The conditions of the deactivation should be connected with an alternative (we change **AND** to **OR**). 11 flats appropriate for the segment (fulfilling all 4 conditions) were found in the search (numbers 10, 12, 17, 35, 88, 101, 105, 122, 130, 132, and 135).

Now we will take into account the features which have a great impact on the client's choice but are not decisive:

- The number of rooms = 3;
- The floor on which the flat is placed = 0;
- The age of the building in which the flat is placed = c. 3 years old;
- Proximity of district A (the time it takes to get to the center) = c. 30 minutes;
- Proximity of a bus or tram stop = c. 80 m.

	Number of rooms	Floor on which the flat is located	Age of the building	Distance of the district center	Proximity of a bus or tram stop
Wanted	3	0	3	30	80
Flat 10	2	1	1	0	150
Flat 12	1	2	1	0	200
Flat 17	3	1	7	20	500
Flat 35	2	0	6	5	100
Flat 88	3	4	6	5	200
Flat 101	4	2	10	0	10
Flat 105	2	2	6	0	50
Flat 122	1	0	6	5	100
Flat 130	2	0	10	0	20
Flat 132	3	5	6	30	400
Flat 135	3	1	6	5	100

Let us note that the last characteristic, i.e. the proximity of a bus or tram stop, is expressed in much greater numbers than the remaining characteristics of the compared flats. As a result that characteristic will have a much greater influence on the obtained result of the distance matrix than the remaining characteristics. In order to prevent it, before the analysis we will normalize all characteristics by choosing a common range for them, from 0 to 1. For that purpose we will use the menu Data→Normalization/Standardization.... In the normalization window we set the "Number of rooms" as the input variable and the empty variable called "Norm(Number of rooms)" as the output variable; the type of the normalization is min/max normalization; the min and max values are calculated from the sample by selecting the button Calculate from sample – the result of the normalization will be returned to the datasheet after selecting the button Run. The normalization is repeated for the following variables, i.e.: "Floor on which the flat is located", "Age of the building", "Distance of the district center", and "Proximity of a bus or tram stop".



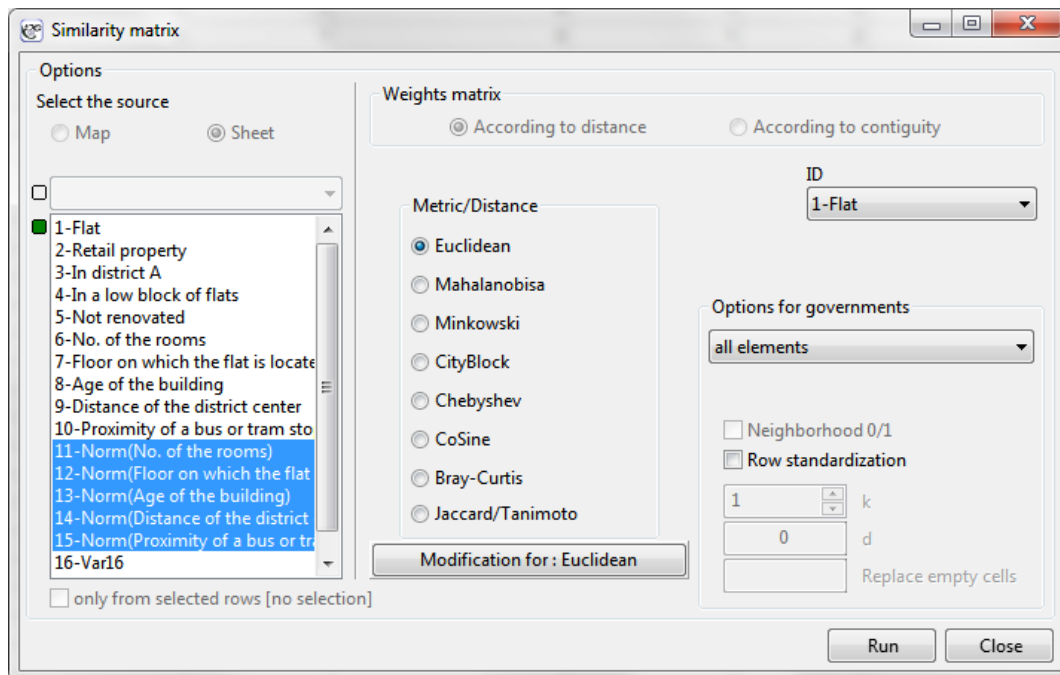
The normalized data are presented in the table below.

	Norm(Number of rooms)	Norm(Floor on which the flat is located)	Norm(Age of the building)	Norm(Distance of the district center)	Norm(Proximity of a bus or tram stop)
Wanted	0,666666667	0	0,222222222	1	0,142857143
Flat 10	0,333333333	0,2	0	0	0,285714286
Flat 12	0	0,4	0	0	0,387755102
Flat 17	0,666666667	0,2	0,666666667	0,666666667	1
Flat 35	0,333333333	0	0,555555556	0,166666667	0,183673469
Flat 88	0,666666667	0,8	0,555555556	0,166666667	0,387755102
Flat 101	1	0,4	1	0	0
Flat 105	0,333333333	0,4	0,555555556	0	0,081632653
Flat 122	0	0	0,555555556	0,166666667	0,183673469
Flat 130	0,333333333	0	1	0	0,020408163
Flat 132	0,666666667	1	0,555555556	1	0,795918367
Flat 135	0,666666667	0,2	0,555555556	0,166666667	0,183673469

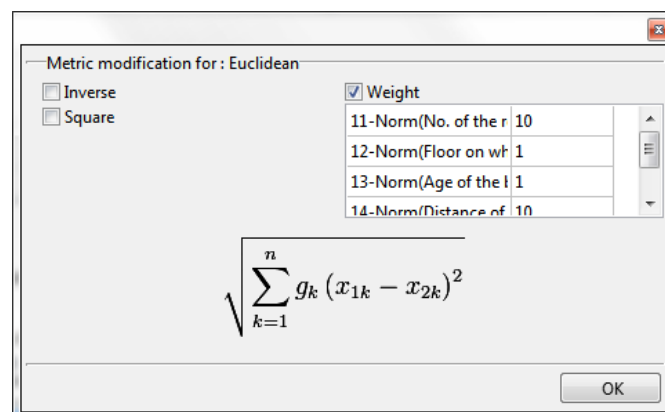
On the basis of the normalized data we will select the flats which are the most suited to the client's inquiry. We will use the Euclidean (distance) metric to calculate the similarity. The smaller the obtained value the more similar the properties. The analysis can be made with the assumption that each of the five characteristics enumerated by the client is equally important but one can also point to the characteristics which should have a greater influence over the result of the analysis. We will build two matrices of Euclidean distances:

- (1) In the first matrix there will be Euclidean distances calculated on the basis of the five characteristics when equally treated;
- (2) In the second matrix there will be those Euclidean distances in the construction of which the number of rooms and the distance to the district center play the most important role.

In order to build the first matrix we select 5 normalized variables in the matrix window, marked as Norm, the Euclidean metric, and the Identifier of the object "Flat" variable.



To build the second matrix we choose, in the matrix window, the same settings as in the case of the first matrix, with the difference that now we additionally select the button Modification: Euclidean and we enter greater weights for the "Number of rooms" and the "Distance of the district center" in the modification window. For example, their values could be equal to 10, and for the remaining characteristics the values could be smaller, e.g. equal to 1.



As a result we will obtain two matrices. In each of them the first column concerns the similarity to the flat looked for by the client:

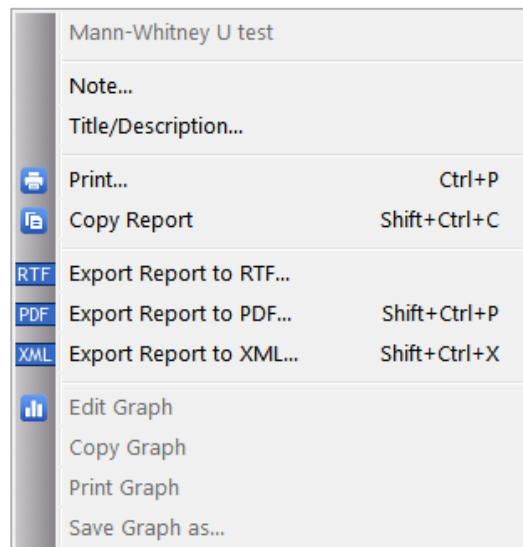
Euclidean	Wanted	...	Weighted euclidean	Wanted	...
Wanted	0	...	Wanted	0	...
Flat 10	1.10	...	Flat 10	3.35	...
Flat 12	1.31	...	Flat 12	3.84	...
Flat 17	1.04	...	Flat 17	1.44	...
Flat 35	0.96	...	Flat 35	2.86	...
Flat 88	1.23	...	Flat 88	2.78	...
Flat 101	1.38	...	Flat 101	3.45	...
Flat 105	1.18	...	Flat 105	3.37	...
Flat 122	1.12	...	Flat 122	3.39	...
Flat 130	1.32	...	Flat 130	3.43	...
Flat 132	1.24	...	Flat 132	1.24	...
Flat 135	0.92	...	Flat 135	2.66	...

According to the unmodified Euclidean distance, the flats best suited to the client's conditions are no. 35 and 135. Having considered the weights, the flats best suited to the client's conditions will be no. 17 and no. 132 which are the first flats with the number of rooms (3) and the distance to the district center similar to that requested by the client. The other 3 characteristics have a smaller influence on the result.

3.2 HOW TO WORK WITH REPORTS (RESULTS SHEETS)

A report is a project element which enables you to store the results of an already done statistic analysis. The report is added automatically to the project and ascribed to the active datasheet at the moment of finishing the current statistic procedure. Note, that it can not be edited, except for graphs and title. Edition of the graph is run by double clicking the mouse or through the context menu of the right mouse button. Title edition is done in the [Project Manager](#) by adding or changing the description.

The main operations of the report can be done via the context menu in the report window




- **Printing**

The options of printing are available by:

- context menu,
- File menu → Print...

- **Export reports**

Reports created in PQStat can be exported to a file in *.rtf (supported by most of text editors such as Word), *.pdf, *.xml.

If the export is made in the [Project Manager](#), the reports can be placed in separate files or in one joint file. To do this, select the adequate reports and then the  button and export to a file or files with the selected format. Individual reports can be exported separately through the context menu in the report window.

- **Describing reports**

Reports can be described in the [Project Manager](#) or in the context menu of report window by adding a title or a note.

- **Editing graphs**

Editing graph relative to its General and Detailed Options is available in the context menu in the report window.

- **Copying reports**

By means of the clipboard, you can also move the results of an analysis into another applications, for example Word or Excel.

- **Deleting reports**

You can delete a report by:



- context menu Delete report (Shift+Del) on the name of the report in the [Navigation tree](#),
- [Project Manager](#).

However, you should remember: if there are any layers of map added to a datasheet and you delete datasheet, all layers attached to it will be deleted too.

The order of reports can be changed with the use of the context menu of the right mouse button Up (Ctrl+Up) or Down (Ctrl+Down) on the name of the report in the [Navigation tree](#).

Adding information to the report name in [Navigation tree](#), such as:

- the hour of generation,
- description,
- filter,
- the name of the grouping variable,
- the name of the variable.

it is possible after selecting an appropriate option in the window of program settings

3.3 HOW TO CHANGE LANGUAGE SETTINGS IN PQSTAT?

Both created reports and program interface can be changed into Polish and English. To change the language, you need to click Edition→Language/Język. Reports opened after the switch, will be translated automatically (except the procedure name, which is the description and is subjected to the user edition).



3.4 MENU

File menu

- New project (Ctrl+N)
- Add datasheet (Ctrl+D)
- Open project (Ctrl+O)
- Open recent
- Open examples
- Import from ...
- Save (Ctrl+S)
- Save as...
- Close project
- Print
- Close (Ctrl+Q) – to close the program

Edit menu

- Undo (Ctrl+Z)
- Cut (Ctrl+X)
- Copy (Ctrl+C)
- Paste (Ctrl+V)
- Delete (Del)
- Select all (Ctrl+A)
- Find/Replace (Ctrl+F)
- Column format (Ctrl+F10)
- Activate/Deactivate (filter)...
- Activate all
- Save selection (Ctrl+T)
- Clear selections
- Language/Język
- Settings

Data menu

- Create table...
- Create raw data...

Sort...

Formulas...

Generate...

Missing data...

Copying with relation...

Normalization/Standardization

Similarity matrix...

Statistics menu

Frequency tables

Descriptive statistics

Probability distribution calculator

- Parametric tests
 - comparison of a one group
 - t-test
 - comparison - dependent groups
 - t-test for dependent groups
 - ANOVA for dependent groups
 - comparison - independent groups
 - t-test for independent groups
 - F Fisher Snedecor
 - ANOVA for independent groups
 - Levene, Brown-Forsythe
 - measures of correlation and their comparisons
 - Linear correlation (r Pearson)
 - Comparison of correlation coefficients
 - measures of agreement
 - ICC - Intraclass Correlation Coefficient
- Nonparametric tests (ordered categories)
 - comparison of a one group
 - Wilcoxon (signed-ranks)
 - Kolmogorov-Smirnov
 - Lilliefors
 - comparison - dependent groups
 - Wilcoxon (matched-pairs)
 - Friedman ANOVA
 - comparison - independent groups

Mann-Whitney
 Chi-square for trend
 Kruskal-Wallis ANOVA

measures of correlation

Monotonic correlation (r Spearman)
 Monotonic correlation (tau Kendall)

measures of agreement

Kendall's W

- Nonparametric tests (unordered categories)

comparison of a one group

Chi-square
 Z for proportion

comparison - dependent groups

Z for 2 dependent proportions
 Bowker-McNemar
 Cochran Q ANOVA

comparison - independent groups

Z for 2 independent proportions
 Chi-square, OR/RR (2x2)
 Fisher, Mid-P (2x2)
 Chi-square (RxC)
 Fisher (RxC)
 Chi-square (multidimensional)

measures of correlation

Q-Yule, Phi (2x2)
 C-Pearson, V-Cramer(RxC)

measures of agreement

Kappa-Cohen

- Diagnostic tests

Diagnostic tests
 ROC Curve
 Dependent ROC Curves – comparison
 Independent ROC Curves – comparison

- Multivariate models

Multiple regression
 Multiple regression - Comparing models
 Logistic regression
 Logistic regression - Comparing models



Principal Component Analysis

Stratified analysis

Mantel—Haenszel OR/RR

- Survival analysis

Life tables

Kaplan-Meier Analysis

Comparison groups

Cox PH regression

Cox PH regression - Comparing models

Scale Reliability

Wizard

Menu Spatial Analysis — description in User Guide - PQStat for Spatial Analysis

Map Manager

Tools

Geometry calculations

Spatial weights matrix

Spatial descriptive statistics

- Spatial Statistics

Nearest Neighbour Analysis

Global Moran's I statistic

Global Geary's C

Local Moran's I statistic

Local Getis-Ord Gi statistic

Menu Graphs

Histogram

Box-Whiskers plot

Error plot

Scatter plot

Line plot

4 HOW TO ORGANISE WORK WITH PQSTAT

All statistic analysis procedures are available in Statistics menu.

4.1 HOW TO ORGANISE DATA

The way of data organisation depends on the statistic procedures, that a user wants to follow.

Statistic analysis of data may be done on the basis of data gathered in a contingency table or as a raw data. But it is also possible to convert data:

- **from a contingency table into a raw form** – you can do this selecting Create raw data... from Data menu,
 - **from a raw form into a contingency table** – you can do this selecting Create table... from Data menu.
1. Data in raw records form are the data organised in the way, so that each row includes information about another studied object (like a patient, a firm etc.).

EXAMPLE 4.1. Raw data (sex-education.pqs file)

<

2. The contingency table presents a joint distribution of 2 variables. There are observed frequencies (natural numbers) inside the table.

EXAMPLE 4.2. A contingency table (sex-education.pqs file)

PQStat v.1.4.0 [C:\Program Files\PQStat\Dane\EN_sex-education.pqs]

File Edit Data Statistics Spatial analysis Help

EN_sex-education

- raw data
- contingency table

[2R x 3C]

	1	2	3	4
	Var1	Var2	Var3	Var4
1		primary+vocational	higher	secondary
2	female	4	7	4
3	male	8	6	5
4				
5				

4.2 HOW TO REDUCE A DATASHEET WORKSPACE

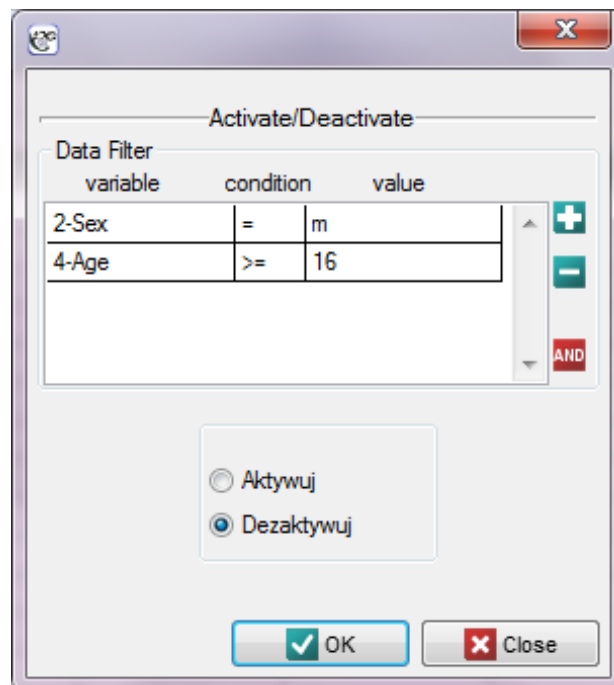
Usually, the whole [datasheet](#) workspace is fully available for you while performing a statistical analysis. However, you can easily limit this area by selecting just a part of the sheet you want to analyse. There are four possible ways to do this:

1. Through activation/deactivation




Activation/deactivation of cases is a global option, superior with respect to other reductions of the area available in the program. Cases (rows) indicated as deactivated are shaded in the data sheet and are not taken into account in statistical analyses.



In order to activate or deactivate selected cases one should choose one of the following options:

- select the rows in the data sheet and choose the option Activate/Deactivate from the context menu on their names;
- select the menu Edit → Activate/Deactivate (filter)...



EXAMPLE 4.3. (file filtr.pqs)

We are going to conduct many statistical analyses on the data from the file filtr.pqs. The analysis will concern boys aged 16 or over. For that purpose we define the rows which will not be analyzed: we select the button  and set the rule for the sex variable; we select the button  again and set the rule for the age variable. Remember: in order to do the exercise correctly all filter conditions should be connected with the conjunction (we are informed about it by the sign ). We set the selected option Deactivate and confirm these analysis conditions by clicking the button OK

When narrowing down the workspace in the data sheet we should remember that the filter conditions can be connected with the conjunction or with the alternative. The change of the alternative and the conjunction is made with the buttons  

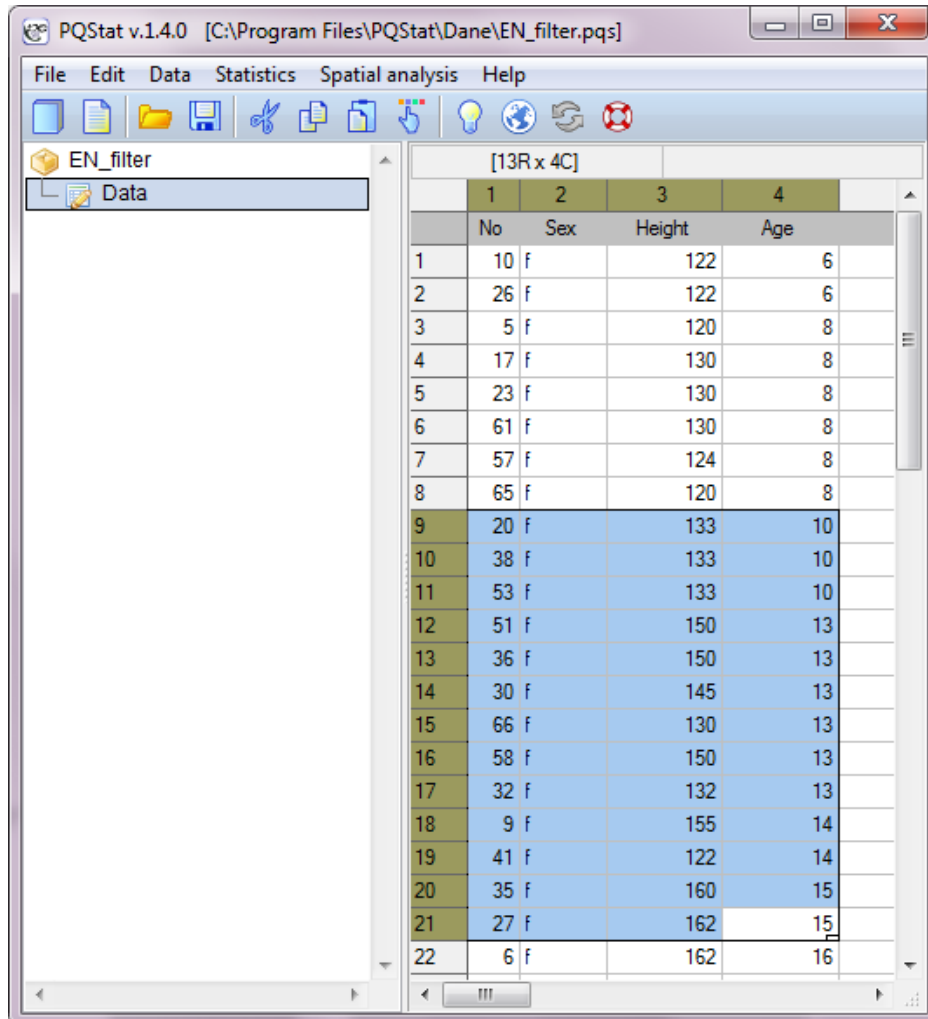
To activate all cases one should select the menu Edit → Activate all

2. You can select the coherent area.

This causes: the analysis we choose is performed using only the selected rows and columns which include necessary data.

EXAMPLE 4.4. (filter.pqs file)

You want to calculate **descriptive statistics** for the height of each girl who is between 10 and 15 years old. In order to calculate this, you need to sort data according to sex and age columns, then you need to select the coherent area of the column which includes 10 to 15 years old girls' height and to select Descriptive statistics from Statistics menu.



	1	2	3	4
	No	Sex	Height	Age
1	10	f	122	6
2	26	f	122	6
3	5	f	120	8
4	17	f	130	8
5	23	f	130	8
6	61	f	130	8
7	57	f	124	8
8	65	f	120	8
9	20	f	133	10
10	38	f	133	10
11	53	f	133	10
12	51	f	150	13
13	36	f	150	13
14	30	f	145	13
15	66	f	130	13
16	58	f	150	13
17	32	f	132	13
18	9	f	155	14
19	41	f	122	14
20	35	f	160	15
21	27	f	162	15
22	6	f	162	16

In the descriptive statistics window, you need to select all procedures that you want to follow (for example mean, standard deviation, minimum, maximum) and the variable for an analysis (the column including height) and then confirm your choice by clicking OK.

If you reduce a datasheet workspace by selecting a coherent piece of data, the following message in the analysed window will occur:

Data reduced by the selected area

3. You can use saved selection.

If selected ranges are ascribed to the sheet, they are highlighted by a frame. They can be used in the analysis, where the data can be set directly to the analysis window. Then, clicking on fill with saved selection button, data from the selected range can be pasted.

EXAMPLE 4.5. (layers.pqs file)

We want to designate statistics associated with Odds Ratio (OR) for a few stratas. We will use some data saved in 10 tables – they are selected (framed). From the Statistics menu, we select Stratified analysis→Mantel-Haenszel OR/RR. In the test options window, we select contingency table, then we set the number of stratas – 10. Each created strata can be filled from the selected range. If we fill all the tables, we make analysis by clicking OK button.

Note



To ascribe more selections to the data sheet from the Edition menu, we chose Save selection (Ctrl+T). To delete ascribed selections, we chose Clear selections.

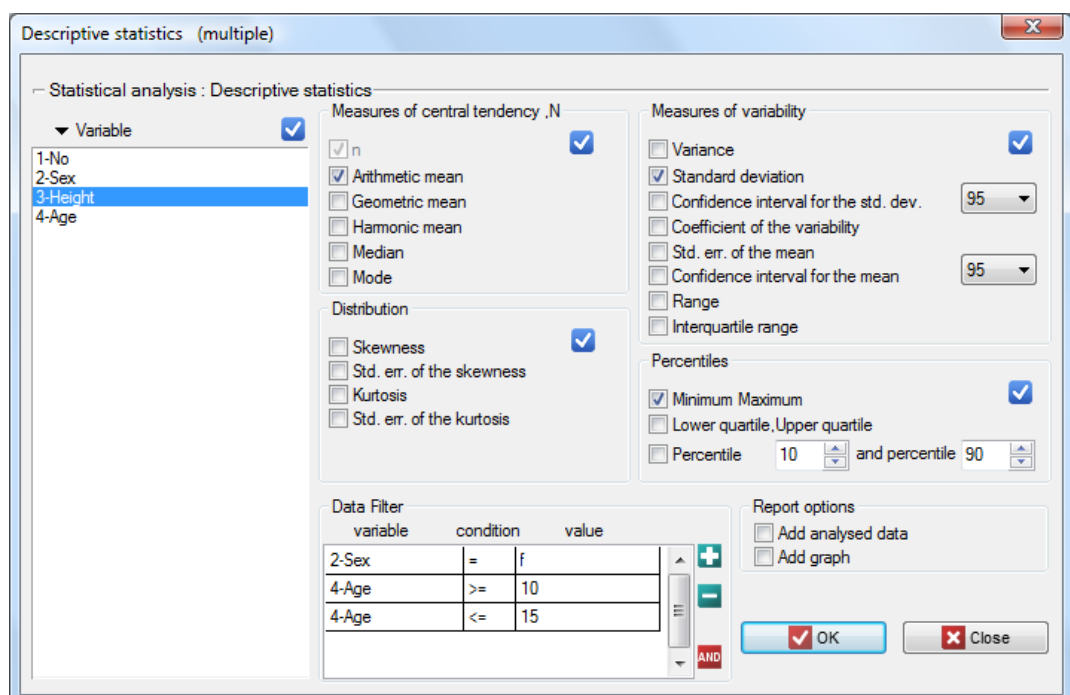
4. You can use a data filter

Data filter is an option which is available when you choose any statistical analysis. If you turn the filter on, the number of rows that are taken into account during the analysis is reduced. There are 2 possible filters: basic filter and multiple filter.

- Basic filter – uses one or more rules joined with conjunctions or alternative.

EXAMPLE 4.6. *Basic filter* (filter.pqs file)

You want to calculate **descriptive statistics** for girls' height, who are between 10 and 15 years old. Choose Descriptive statistics from Statistics menu. In the descriptive statistics' options window, you should select all the procedures you want to have done (for example you select mean, standard deviation, minimum and maximum) and variable for analysis (column which includes height). To set filter you need to add rules using  button. First, you need to set the rule for the variable - sex. Then, choose "equal" sign as a condition and "g" letter, which means girls, as a value. After that, you should add another rule and set the the variable - age. Then, \geq sign as a condition and 10 as a value. Exactly the same way you add age condition ≤ 15 . Note, to do this task properly, all the rules of the filter should be joined with conjunction (the  sign informs you about it). If you select analysis conditions properly, confirm your choice by clicking OK.



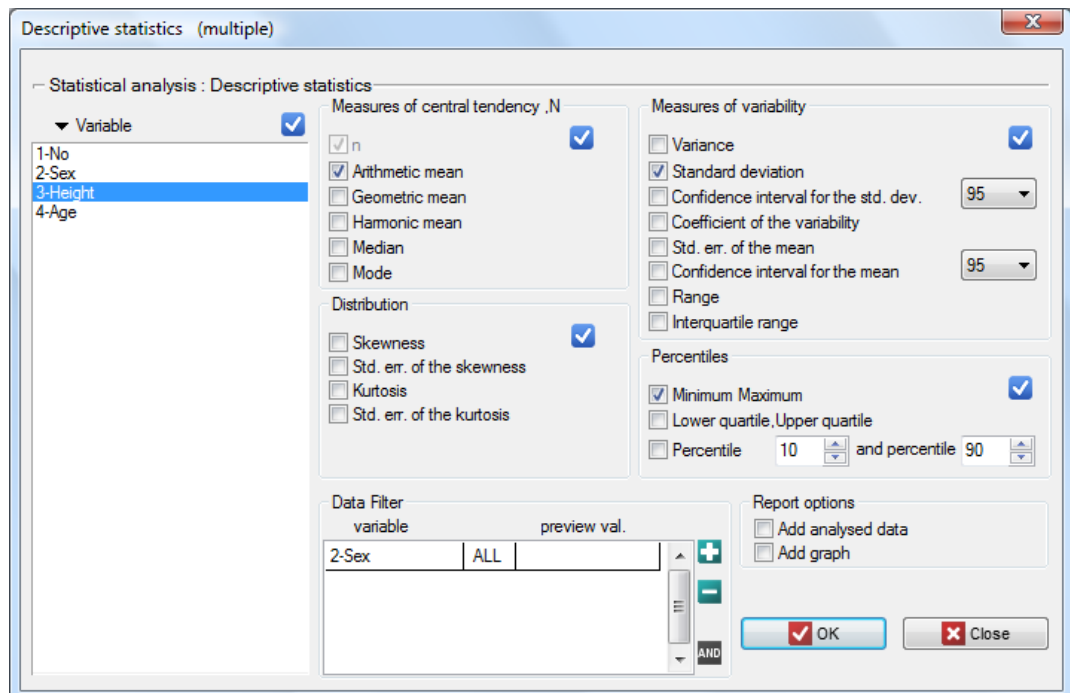
Remember, when reducing a datasheet workspace using a data filter, filter conditions may be matched with a conjunction or an alternative. To change alternative and conjunctions,

use **AND** **OR** buttons.

- Multiple filter – uses one rule to divide data into several subgroups. The selected analysis is performed several times, separately for each subgroup.


EXAMPLE 4.7. Multiple filter (filter.pqs file)

You want to calculate **descriptive statistics** for girls' height and for boys' height separately. Choose Descriptive statistics from Statistic menu. In the option window of descriptive statistics choose procedures you want to have done (select for example mean, standard deviation, minimum and maximum) and variable to make analysis (column including age). Select multiple filter and add rule using **+** button. As a rule select the variable - sex. At the end, confirm all chosen options by clicking OK. As a result you get 2 reports: separately for boys and separately for girls.



4.3 MULTIPLE REPEATED ANALYSIS

To improve the performance of repeated analyses, you can:

1. Use the option of saving current analysis. PQStat program saves recently performed analysis and its settings. To go back to this analysis quickly, just click  button on the toolbar.
2. In the analysis window, choose many variables so that the analysis will be carried out repeatedly. Results of the analyses will be returned in the following reports.
3. Use the multiple [filter](#) so that the analysis will be carried out separately for individual subsets of data. Results of the analyses will be returned in the following reports.

4.4 INFORMATION GIVEN IN A REPORT

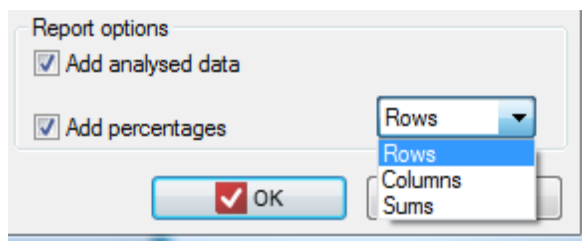
Apart from basic settings, which refer to the already done statistic analysis, in the test window, there is a possibility to:

- Add analysed data to a report.
Analysed data, depending on the test, are given to the report:

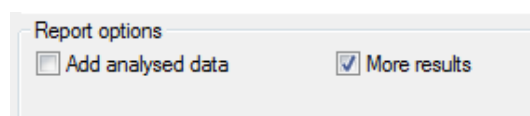
- as a [raw data](#),
- as a [contingency table](#).

Additionally, it is possible to view contingency table of proportional values calculated from:

- table row,
- table column,
- total sum of the table.



- Add graph to a report.
To add an appropriate graph to the report, select option Add graph in the window of a particular statistical analysis.
- Limitations of numbers of returned results.
If there are any statistical tests whose reports include a lot of results, you can limit the amount of returned information by deselecting the option Full calculations:



4.5 MARKING OF STATISTICALLY SIGNIFICANT RESULTS

In the report, a p -value of performed statistical test is marked with red colour only if the [p value](#) is less than a [significance level](#) defined by the user. The default significance level for all tests is = 0.05. You can change this setting permanently in the [Settings](#) window or just temporarily (till the application is opened) in the window of the chosen test.



5 GRAPHS

The PQStat program offers column charts, error charts, box plots, point charts, and line and point charts.

The window with the settings of the the options of graphs is called up via the menu Graphs.

The change of the basic parameters of the graph is possible directly in the graph window. If:

- we want to change the general graph parameters, such as: titles, backgrounds, axes, grid lines, or the legend – we choose the tab Graph General Options;
- we want to change the appearance of the drawn object, e.g. the shape, style, colors – we choose the tab Graph Detailed Options;
- we want to draw additional elements e.g. line – we choose the tab Others.

The graphs presenting the results of statistical analyses are available in the window of the selected statistical analysis at the option Add graph.

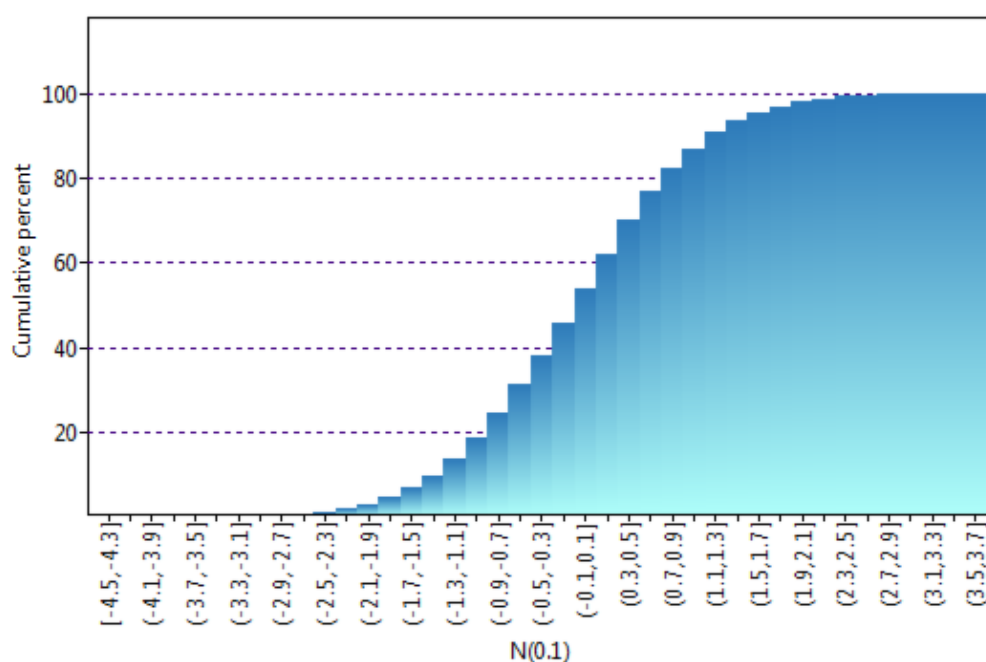
The graph is returned to the report where it can be:

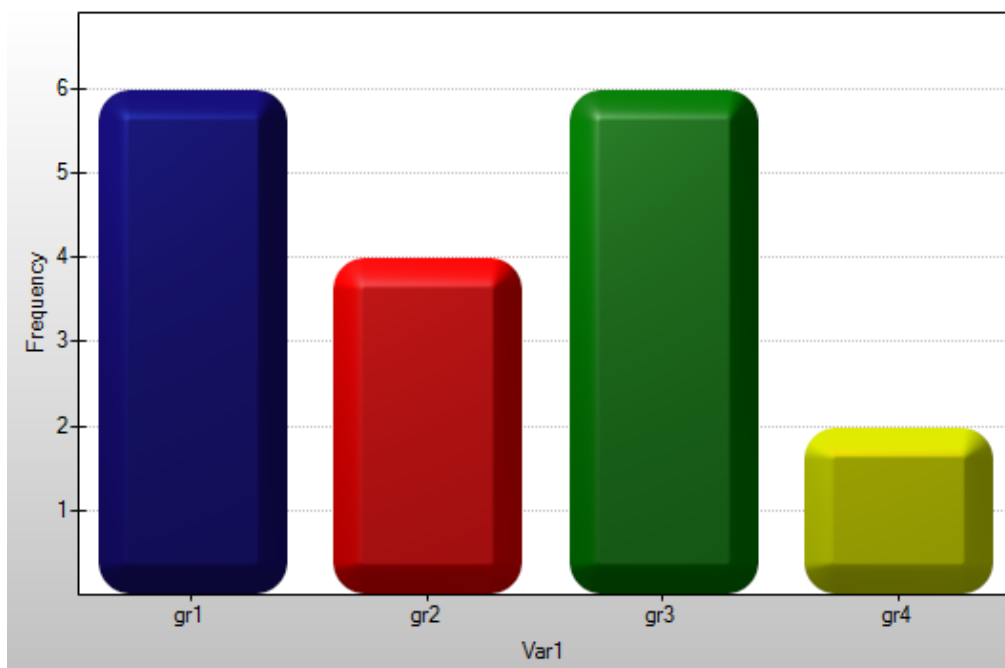
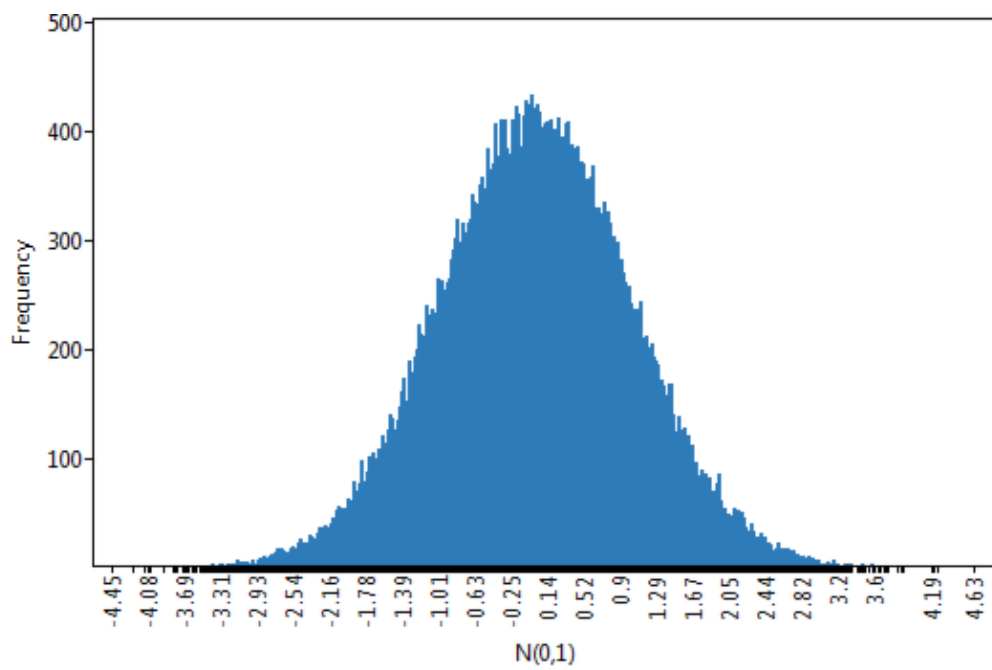
- saved – option Save Graph as... from the context menu;
- printed – option Print Graph from the context menu;
- copied – option Copy Graph from the context menu;
- edited – this applies to the Graph General Options and Graph Detailed Options. To edit a graph it is enough to double-click on the graph or to choose the option Edit Graph from the context menu. In the edition window it is also possible to save the graph at high resolution.

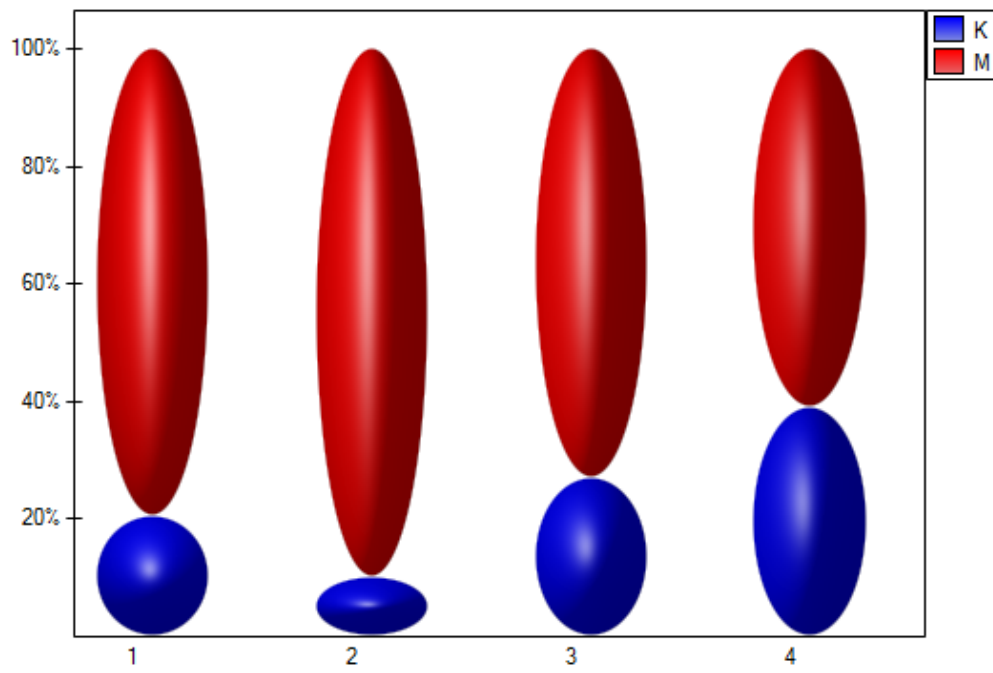
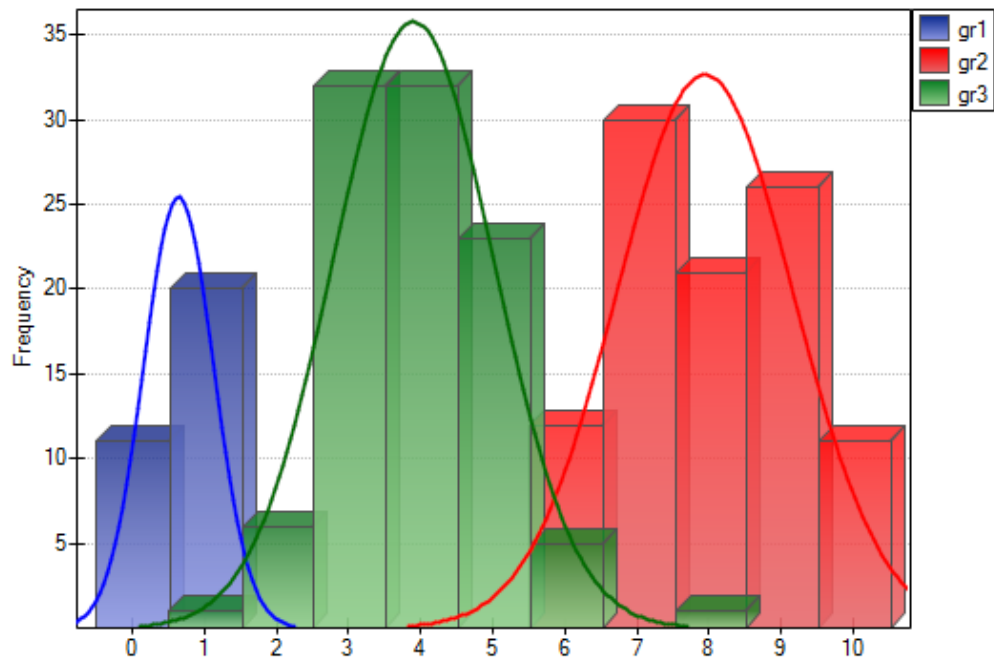
5.1 GRAPHS GALLERY

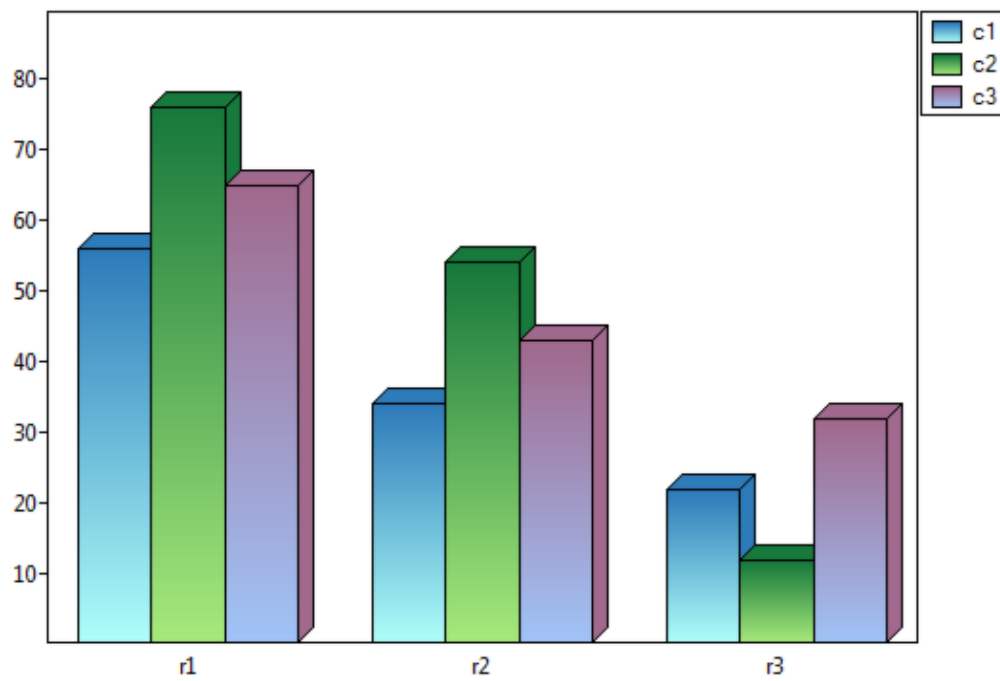
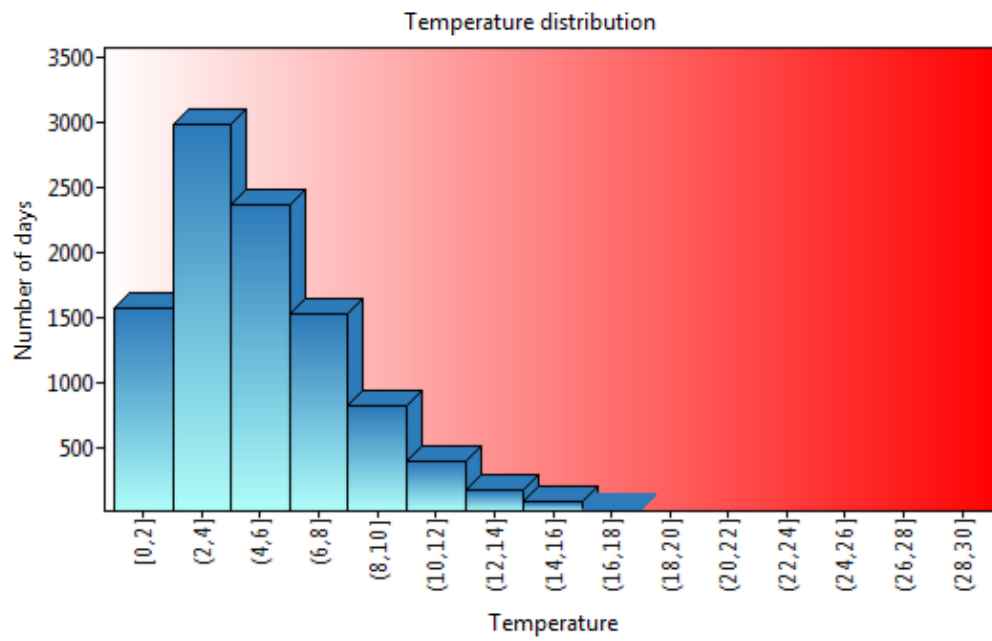
According to the type of analysis, there is a various choice of graphs:

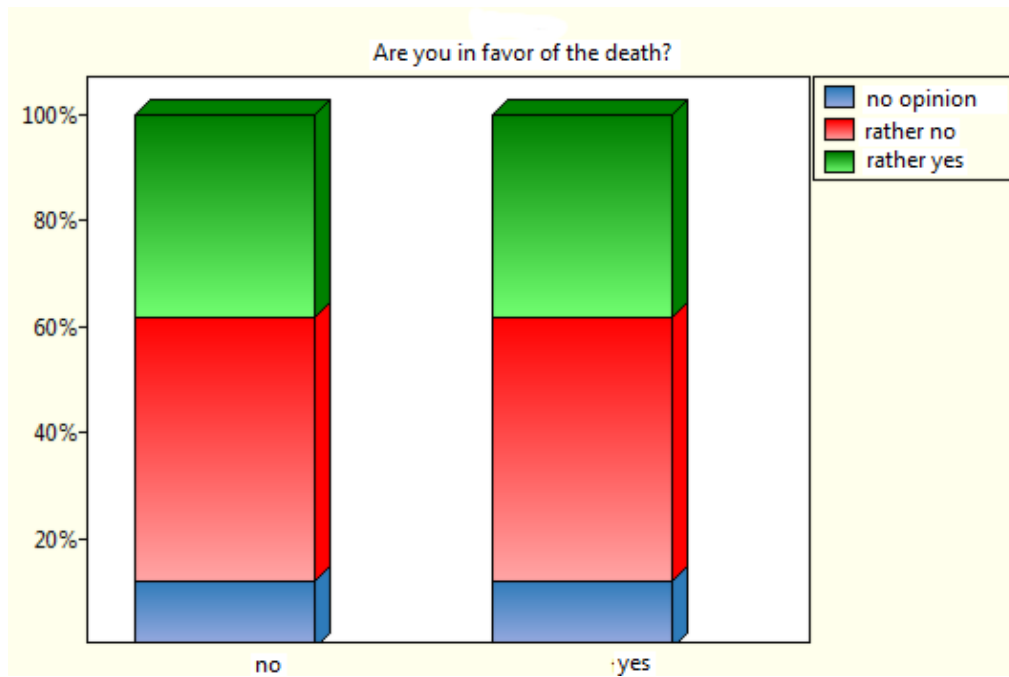
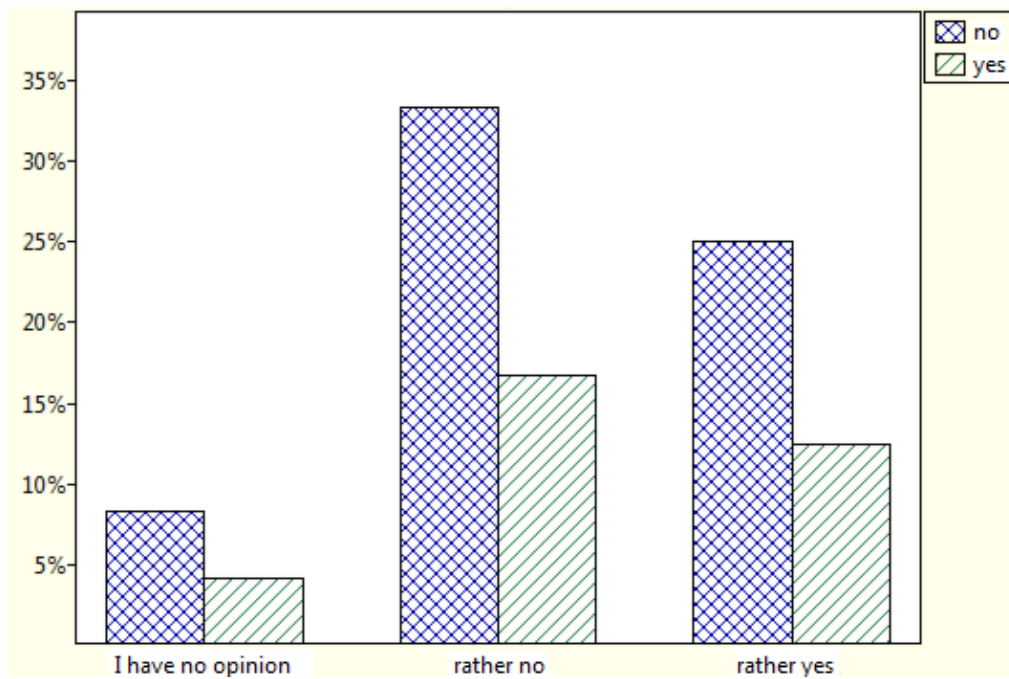
5.1.1 Bar plots

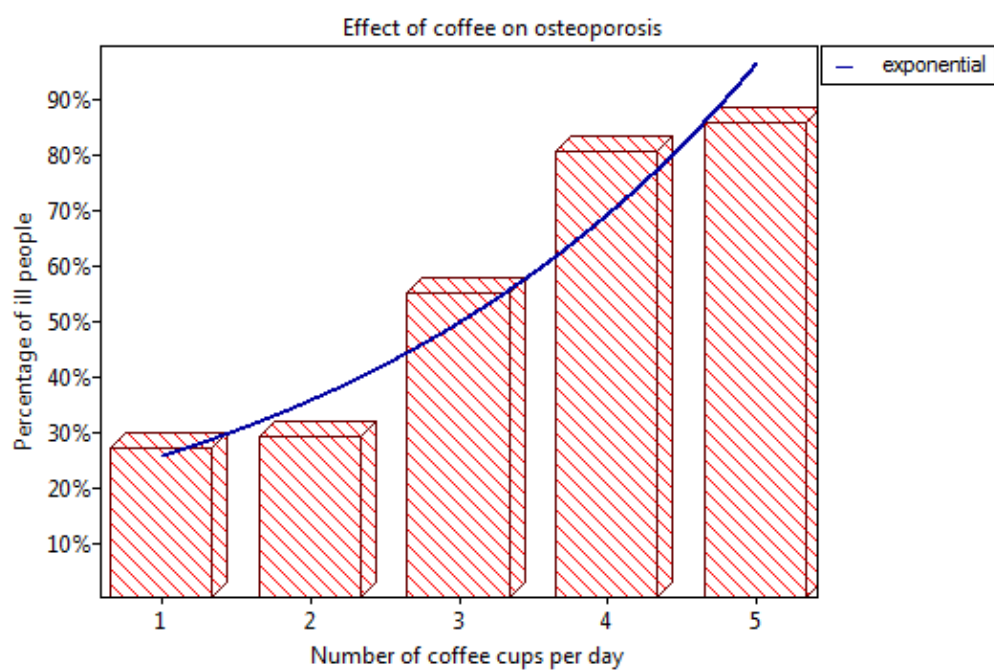




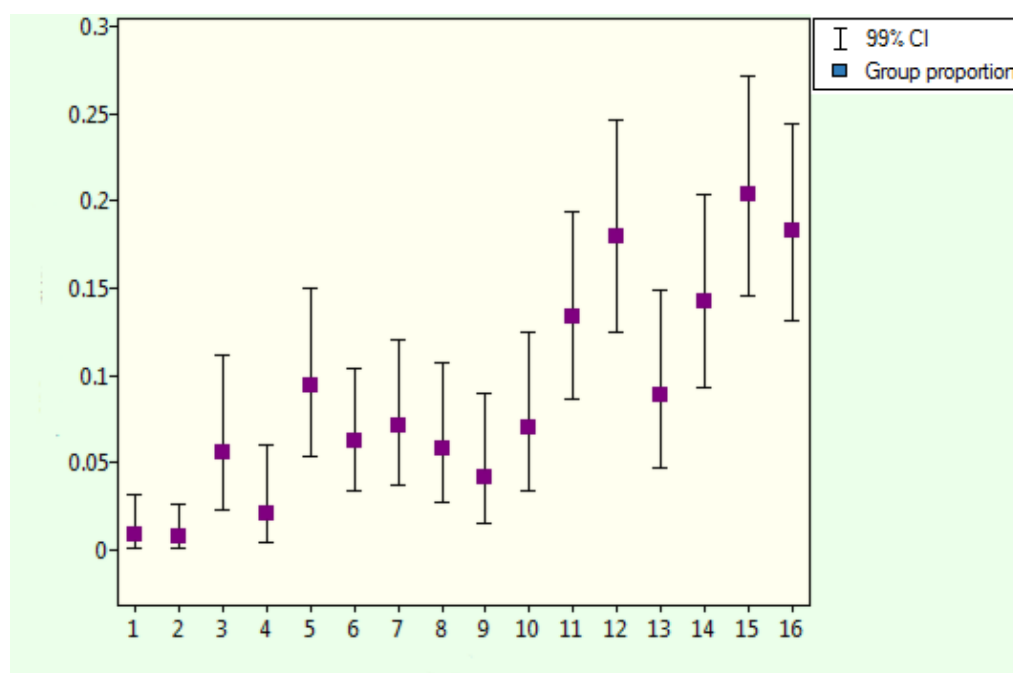


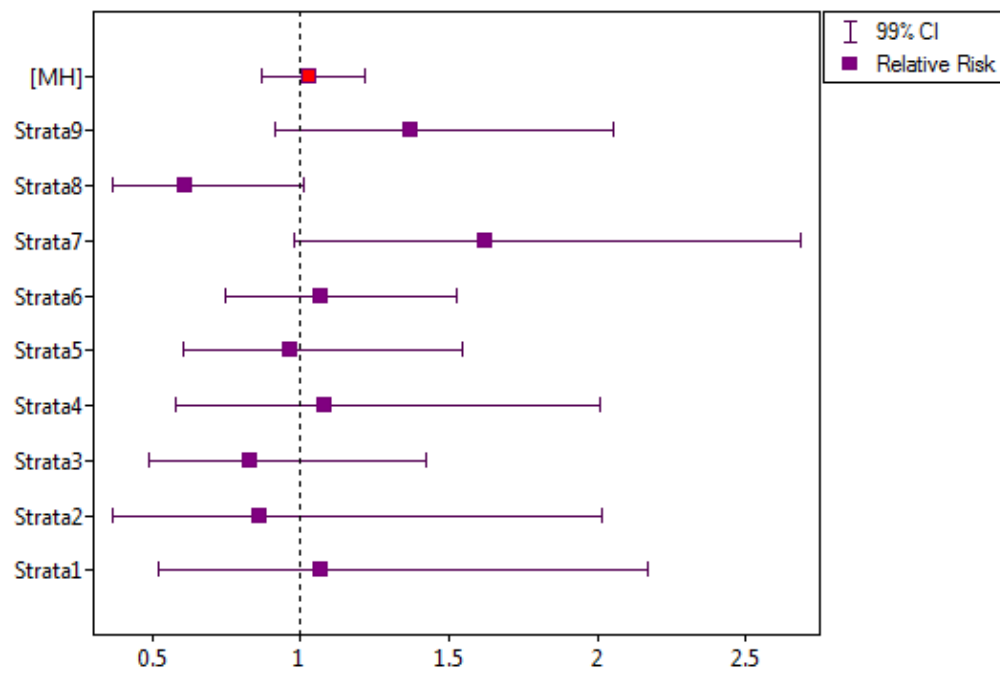
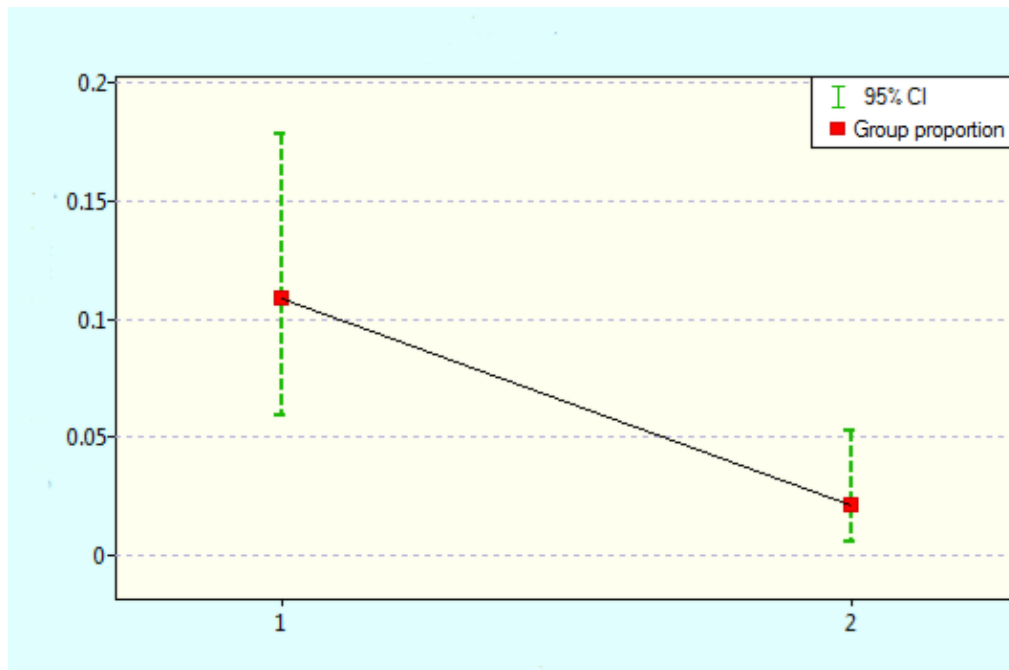




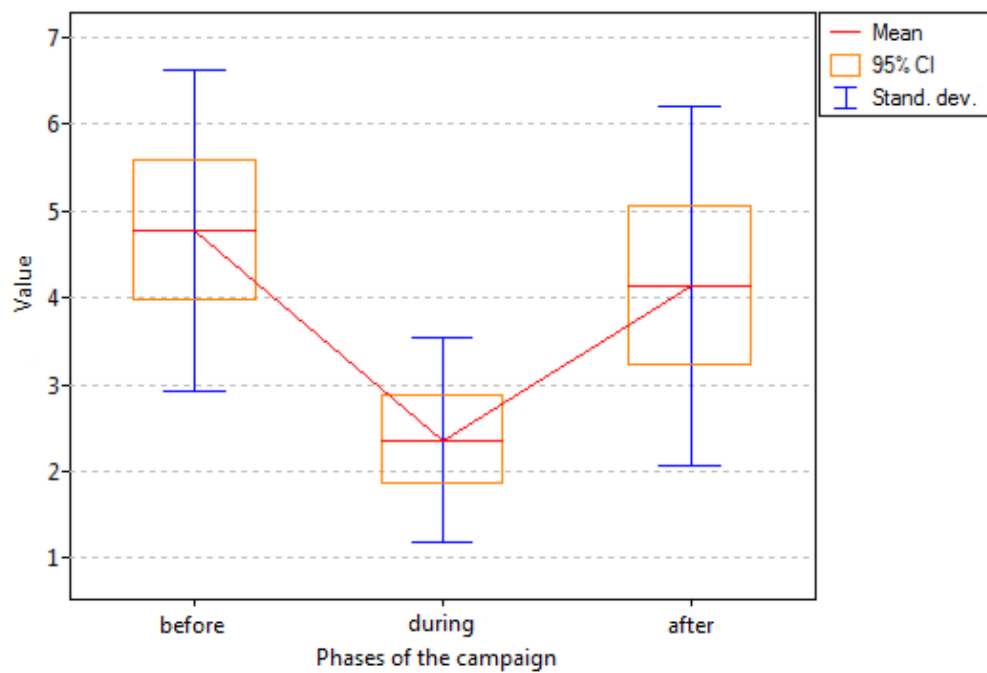
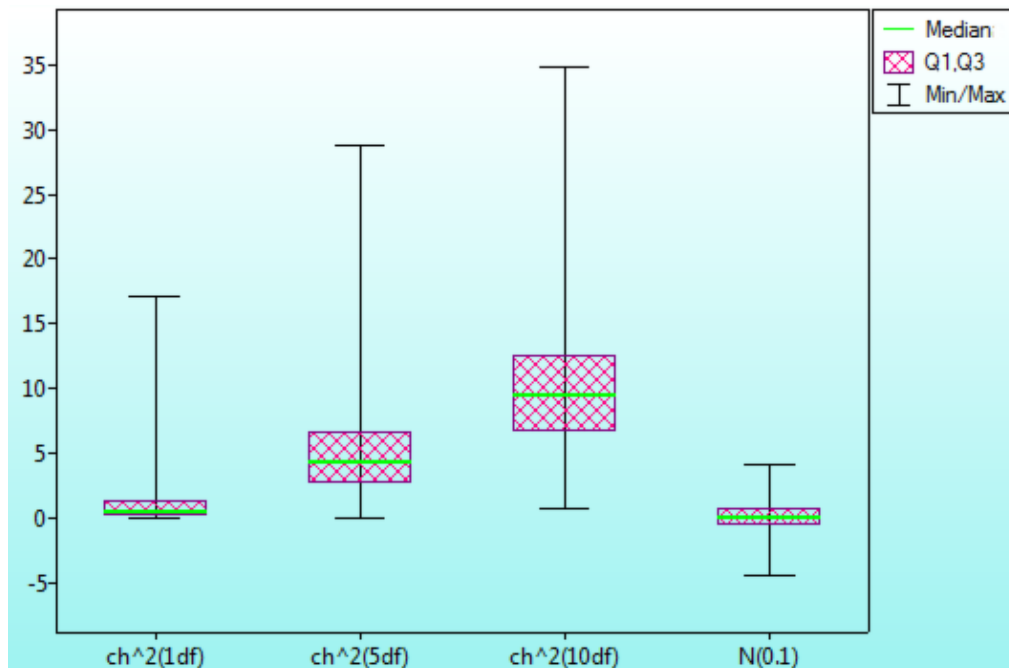


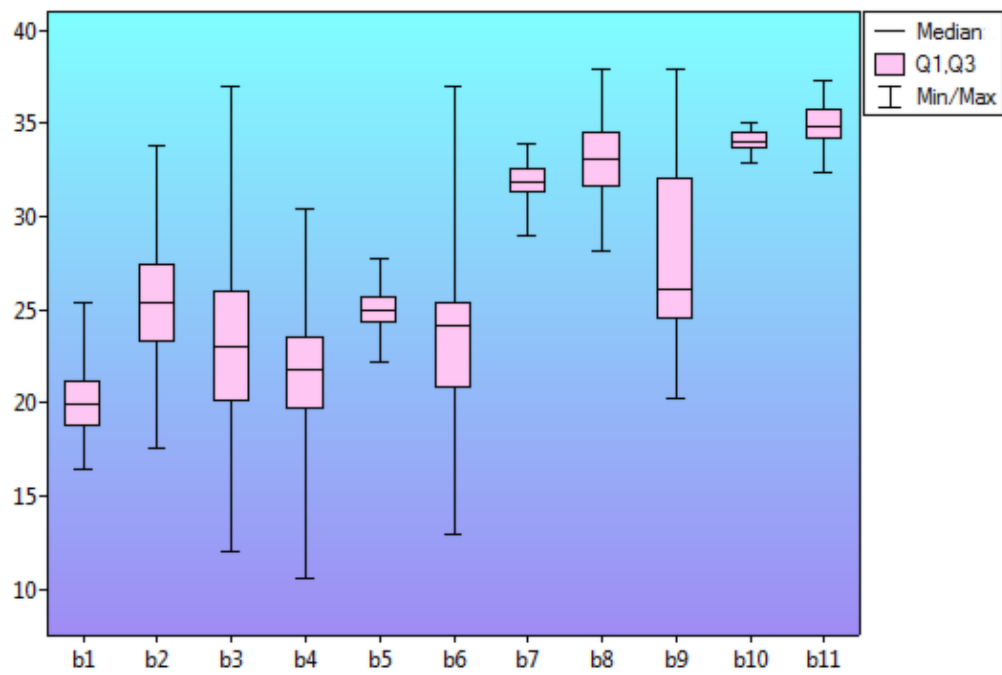
5.1.2 Error plots



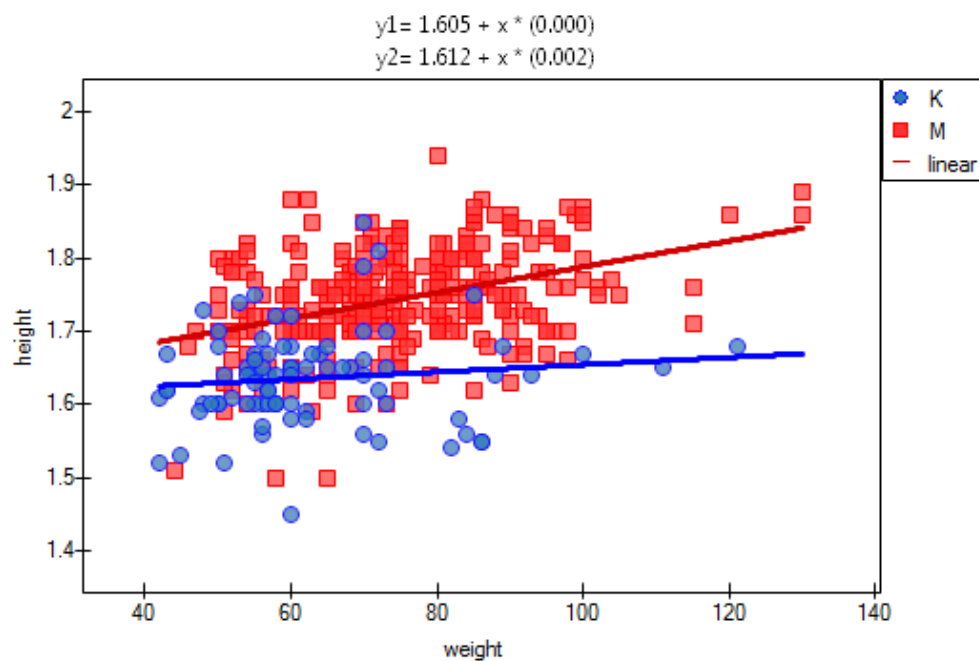


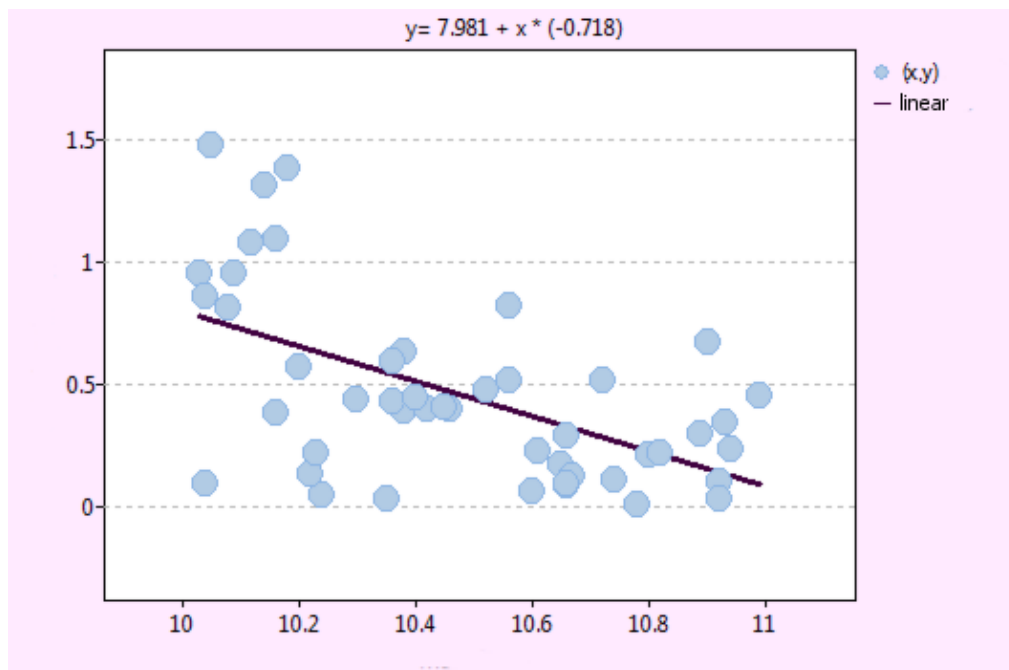
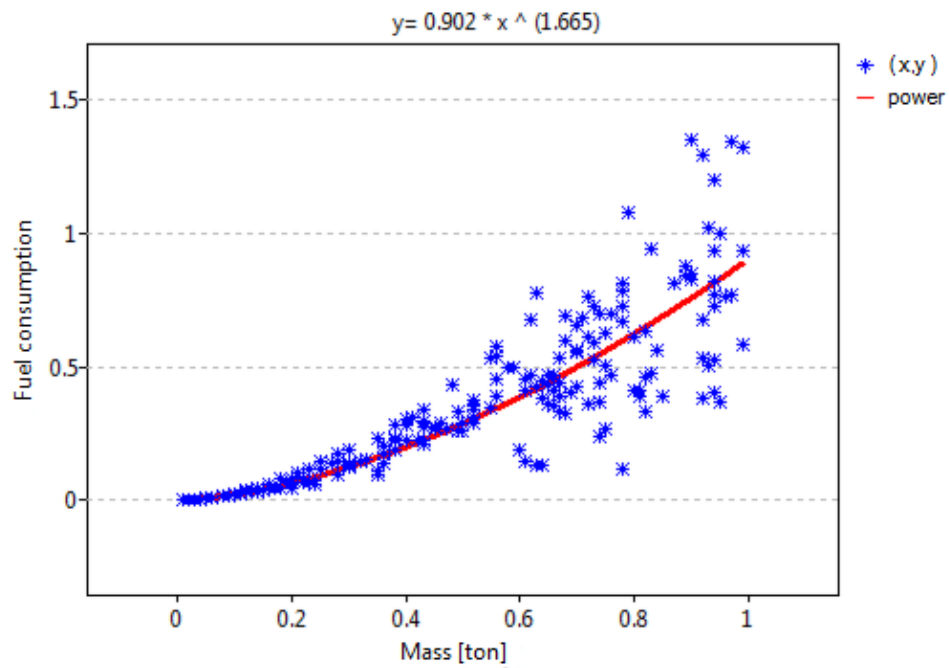
5.1.3 Box-Whiskers plots

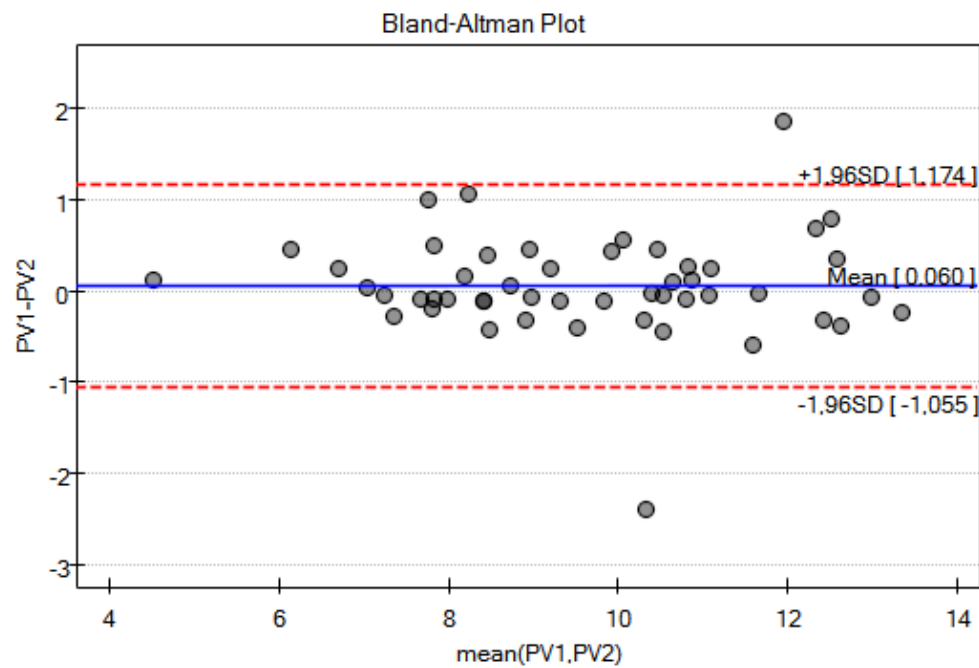




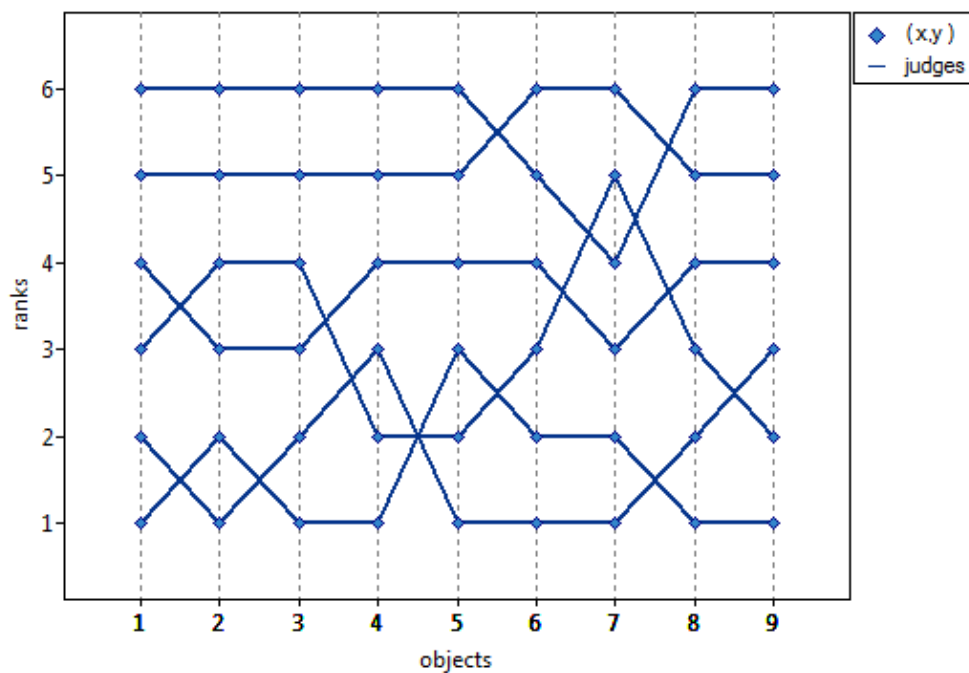
5.1.4 Scatter plots

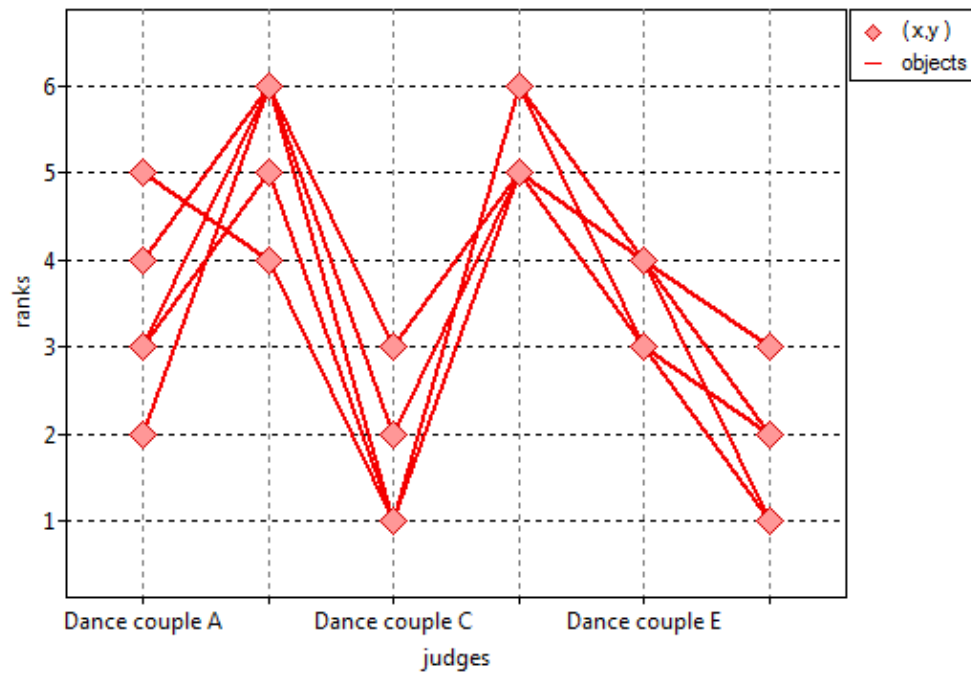






5.1.5 Line plots

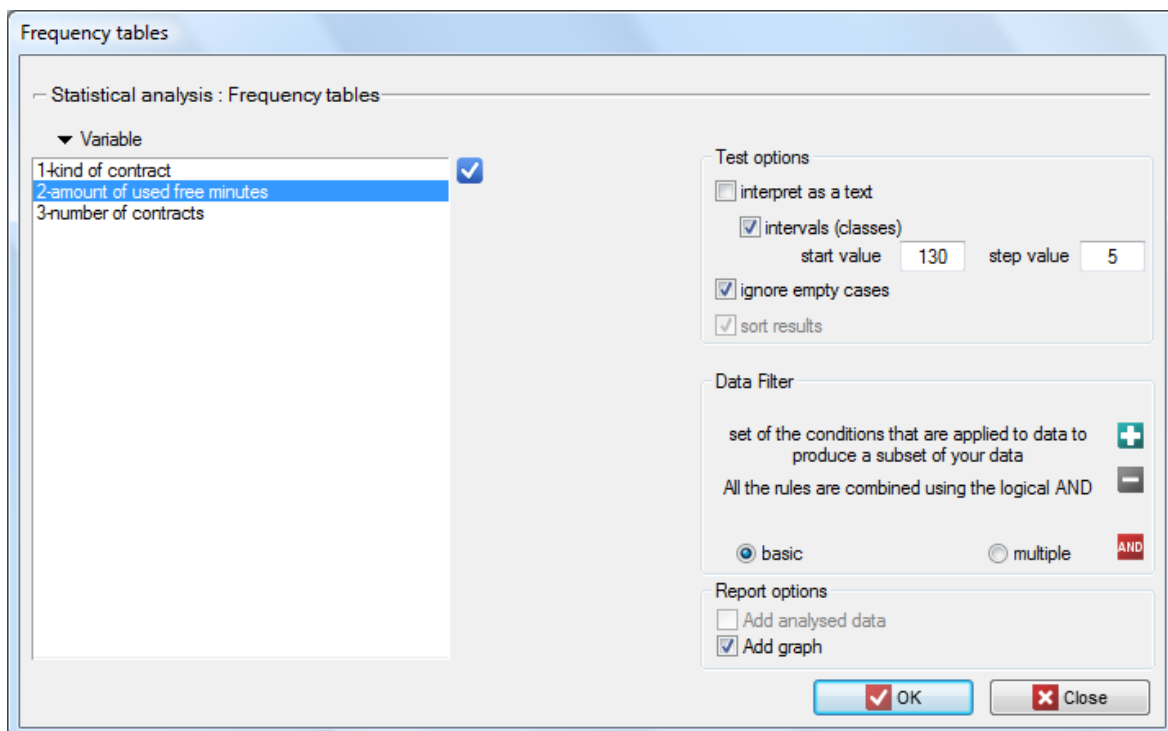




6 FREQUENCY TABLES AND EMPIRICAL DATA DISTRIBUTION

The basis of all statistical analyses is to define an **empirical distribution**, in other words - the observed feature distribution in a sample. To define an empirical feature distribution, you need to assign the frequency of occurrence to the following values of this feature. Such distribution may be presented either in a **frequency tables** or in a graph (histogram). For small data sets, the frequency table can show all the data - so called a frequency distribution. For the larger data sets they are called a grouped frequency distribution.

To present data distribution in a table, you need to display Frequency tables window by selecting Statistics menu→Frequency tables.



In this window, you should select a variable that you want to have analysed and analysis options. If the options are chosen properly, we can sort the calculated result treating variables as text values or numbers. If there are empty cells in an analysed column, they can be included or omitted in an analysis. The result of a particular analysis will occur in a report added to a datasheet, for which the analysis have been done.

Additionally, if we want the data to be illustrated in a bar plot or a histogram, we select Add graph option in the Frequency tables.

EXAMPLE 6.1. (distribution.pqs file)

Some mobile network operator did the research, which was supposed to show the use of "free minutes" given to his clients on a pay-monthly contract. Each customer may use up to 190 free minutes every month. The research was done on the basis of 200 clients. There were several sorts of information taken into account:

- the kind of contract,
- the amount of used free minutes,
- the number of contracts taken by one client (it does not apply to companies).

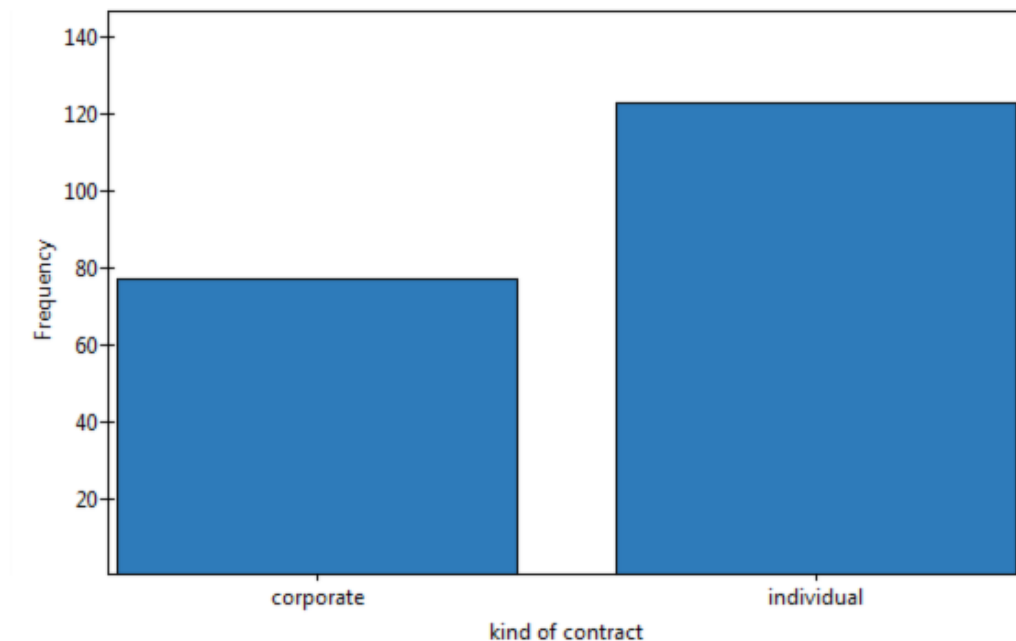
Now you want to present distribution of:


1. the kind of contract,
2. the amount of used free minutes,
3. the number of registered contracts with individual persons.

Open the Frequency tables window.

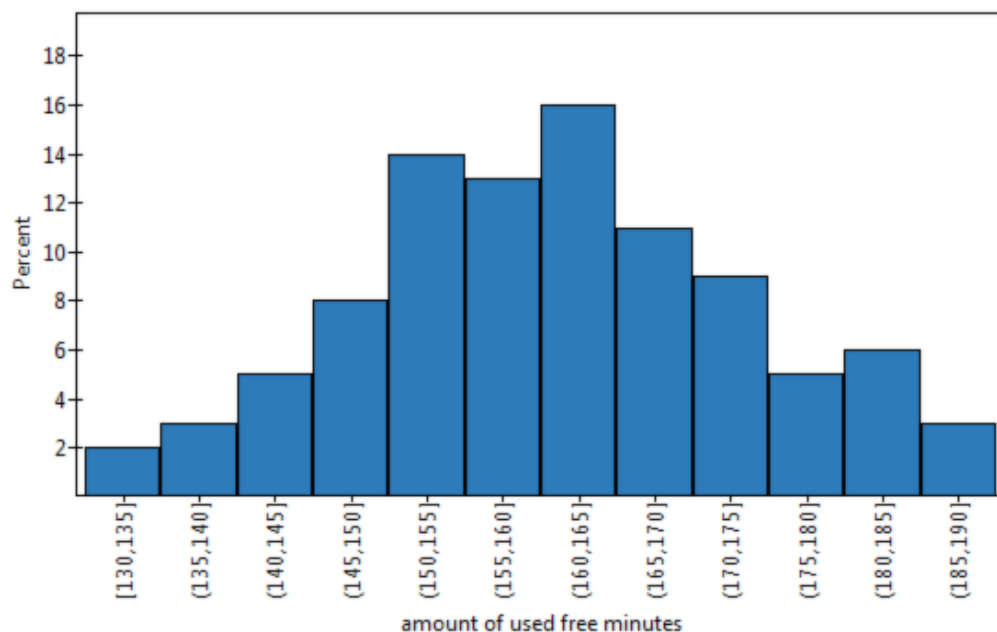
1. Choose the variable that you want to analyse: "the kind of contract" and select the option to interpret it as a text value and Add graph. Then confirm all the chosen settings by clicking OK and you get the result presented in a report:


Frequency tables				
Analysis time	0.09sec.			
Variable: kind of contract	Frequency	Cumulative frequency	Percent	Cumulative percent
corporate	77	77	38.5%	38.5%
individual	123	200	61.5%	100%



2. Do the **analysis again** by clicking  button. Choose the variable that you want to analyse: "the amount of used free minutes" and then the option Intervals (ranks), set start value, which is for example 130 and a step value is 5. You may also select Add graph option. Next, confirm all the chosen options by clicking OK and you get the result presented in a report:

Frequency tables					
Analysis time		0.02sec.			
Variable: amount of used free minutes	Frequency	Cumulative frequency	Percent	Cumulative percent	
[130,135]	5	5	2.5%	2.5%	
(135,140]	7	12	3.5%	6%	
(140,145]	11	23	5.5%	11.5%	
(145,150]	17	40	8.5%	20%	
(150,155]	29	69	14.5%	34.5%	
(155,160]	27	96	13.5%	48%	
(160,165]	32	128	16%	64%	
(165,170]	23	151	11.5%	75.5%	
(170,175]	19	170	9.5%	85%	
(175,180]	11	181	5.5%	90.5%	
(180,185]	13	194	6.5%	97%	
(185,190]	6	200	3%	100%	



- Do the [analysis again](#) by clicking the  button. Set the [filter](#) so that the analysis is carried out only for individual persons. Choose the variable you want to analyse: "the number of contracts". This variable includes missing data (empty cases), that is why they may be taken into account as well as not be taken in the result. It depends on the chosen option which refers to ignoring (or not) the empty cases:

Frequency tables				
Analysis time		0.01sec.		
Data Filter:				
kind of contract=individual				
Variable: number of contracts	Frequency	Cumulative frequency	Percent	Cumulative percent
1	94	94	79.661%	79.661%
2	12	106	10.169%	89.831%
3	11	117	9.322%	99.153%
4	1	118	0.847%	100%

Frequency tables				
Analysis time		0.01sec.		
Data Filter:				
kind of contract=individual				
Variable: number of contracts	Frequency	Cumulative frequency	Percent	Cumulative percent
1	94	94	76.423%	76.423%
2	12	106	9.756%	86.179%
3	11	117	8.943%	95.122%
4	1	118	0.813%	95.935%
empty	5	123	4.065%	100%

EXAMPLE 6.2. (fertiliser.pqs file)

There was made an experiment in order to analyse a microbiological condition of the soil, where the fertilised (with biologically active fertilisers) perennial ryegrass is grown. The soil was fertilised with various microbiological specimen and fertilisers. After that, there was a number of microorganisms which occurred in a 1 gram of dry mass of calculated soil. You want to get to know the frequency of actinomycetes occurrence in a 1 gram of dry mass of the soil fertilised with nitrogen. You want to find out how often, in the analysed sample, values of actinomycetes had occurred (in intervals: from 0 to 20 , from over 20 to 40, from over 40 to 60, ...). You need to select only the 54 first rows in a datasheet, which fulfil the analysis Assumptions (there are actinomycetes fertilised with nitrogen) and then to open a frequency tables window in Statistic menu→Frequency tables.

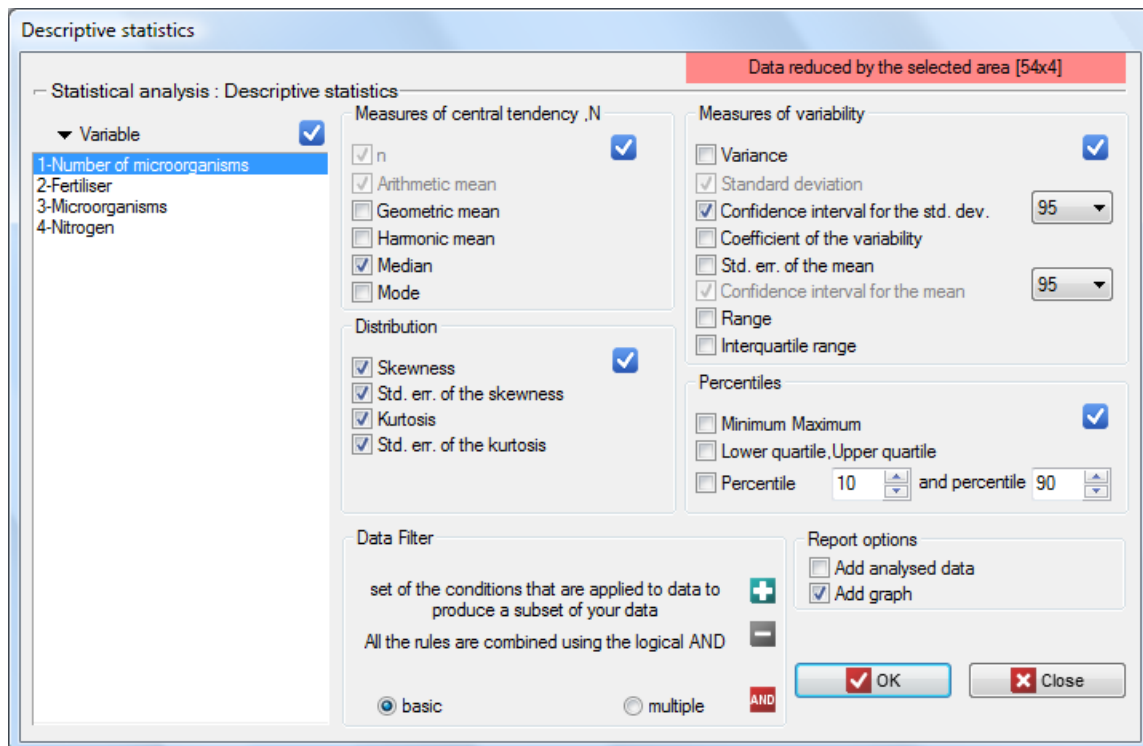
In the options window, you need to select a variable which you want to analyse: The number of microorganisms. After that you need to set ranges (classes), so that the start value is 0 and the step value is 20. At the top of the window you should see the message: **Data reduced by the selected area**. Now confirm your choice by clicking OK and you will get a result presented in the report.

Frequency tables				
Analysis time		0.09sec.		
Variable: Number of microorganisms	Frequency	Cumulative frequency	Percent	Cumulative percent
[0,20]	1	1	1.852%	1.852%
(20,40]	3	4	5.556%	7.407%
(40,60]	6	10	11.111%	18.519%
(60,80]	21	31	38.889%	57.407%
(80,100]	16	47	29.63%	87.037%
(100,120]	4	51	7.407%	94.444%
(120,140]	3	54	5.556%	100%

7 DESCRIPTIVE STATISTICS

We use descriptive statistics to describe main features of the collection of data, for example mean value, median or standard deviation and to draw some basic conclusions and generalisation about the collection of data.

To calculate descriptive statistics for data gathered in a sheet, you should open the Descriptive statistics window which is in Statistics menu→Descriptive statistics.



In this window, you need to select variables you want to analyse and then select all the descriptive statistics measures you need for the analysis. However, note that you can select separate statistics or groups of statistics using ☒ button. Confirm your choice by clicking OK. The result of the analysis will be presented in a report added to the datasheet, on the basis of which the analysis was done.

Additionally, if we want the data to be illustrated in a Box-Whiskers plot, we select Add graph option in the Descriptive statistics window.

7.1 MEASUREMENT SCALES

A properly defined kind of an analysis depends on the scale, on which the data are presented. There are 3 main measurement scales:

1. Interval scale

Variables are assessed on an interval scale if:

- it is possible to order them,
- it is possible to calculate how much one element is greater than the other one and the differences between these elements are interpretable in a real world. Usually the unit of measurement is defined.

Example: the mass of an object [kg], the area of an object [m], time [years], speed[km/h] etc.

2. Ordinal scale

Variables are assessed on an ordinal scale if:

- it is possible to order them, so the sequence of occurred elements does matter,
- it is impossible to define the quotient and the difference between two values in a logical way.

Example: education, competitors order on the podium, etc.

Note

Note that, if a variable is assessed on an ordinal scale, to enable proper calculations on it, it should be written by means of numbers. Numbers are a kind of agreed identifiers telling us about the order of elements.

3. Nominal scale

Variables are assessed on a nominal scale if:

- it is impossible to order them, because there is no order resulting from the nature of the given occurrence,
- it is impossible to define the quotient and the difference between two values in a logical way.

Example: sex, country of residence etc.

Note

If a variable is assessed on a nominal scale, it can be written by means of text labels. Even if the values of a nominal variable are written in numbers, these numbers are only a kind of agreed identifiers, so it is impossible to make any arithmetical calculations based on them and it is also impossible to compare them.

7.2 MEASURES OF POSITION (LOCATION)

7.2.1 CENTRAL TENDENCY MEASURES

Central tendency measures are so called average or mean measures whose characteristic is mean or a typical level of a feature value.

Arithmetic mean is formulated:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n},$$

where x_i means following values of variable and n means a sample size.

Arithmetic mean is used for an [interval scale](#). If used for a sample, it should be marked with \bar{x} , but for population with μ .

Geometric mean is formulated:

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \dots x_n} = \sqrt[n]{\prod_{i=1}^n x_i}.$$

This mean is used for an [interval scale](#) if the variable distribution is log-normal, so the variable logarithm has a [normal distribution](#).

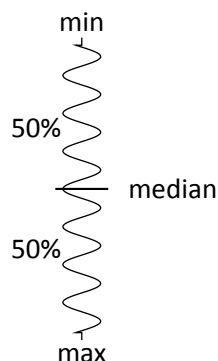
Harmonic mean is formulated:

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}.$$

This mean is used for an [interval scale](#).

Median

In the ordered data set, median is the value that divides this set into two equal parts. Half of all observations is below and half of them is above the median.



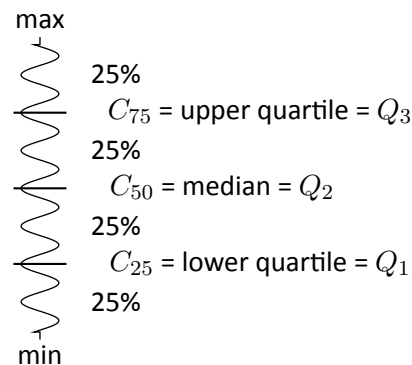
Median can be used in both [interval](#) and [ordinal scale](#).

Mode

Mode is a value that occurs the most often among the results. Mode can be used in each [measurement scale](#).

7.2.2 ANOTHER MEASURES OF POSITION

Quartiles, deciles, centiles



Quartiles (Q_1, Q_2, Q_3) divide an ordered rank into 4 equal parts, deciles ($D_i, i = 1, 2, \dots, 9$) divide an ordered rank into 10 equal parts and centiles (percentiles: $C_i, i = 1, 2, \dots, 99$) into 100 equal parts. The second quartile, the fifth decile and the fiftieth centile are equal to median. These measures can be used in an [interval](#) or [ordinal scale](#).

7.3 MEASURES OF VARIABILITY (DISPERSION)

Central tendency measures knowledge is not enough to fully describe a statistical data collection structure. The researched groups may have various variation levels of a feature you want to analyse. You need some formulas then, which enable you to calculate values of variability of the features.

Measures of variability are calculated only for an **interval scale**, because they are based on the distance between the points.

Range is formulated:

$$I = \max x_i - \min x_i,$$

where x_i are values of the analysed variable

$$IQR = \text{Interquartile range} = Q_3 - Q_1,$$

where Q_1, Q_3 are the lower and the upper quartile.

Ranges for a percentile scale (decile, centile)

Ranges between percentiles are one of the dispersion measures. They define a percentage of all observations, which are located between the chosen percentiles.

Variance — measures a degree of spread of the measurements around arithmetic mean

sample variance:

$$sd^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1},$$

where x_i are following values of variable and \bar{x} is an arithmetic mean of these values,
n - sample size;

population variance:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N},$$

where x_i are following values of variables and μ is an arithmetic mean of these values,
N - population size;

Variance is always positive, but it is not expressed in the same units as measuring results.

Standard deviation — measures a degree of spread of the measurements around arithmetic mean.

sample standard deviation:

$$sd = \sqrt{sd^2},$$

population standard deviation:

$$\sigma = \sqrt{\sigma^2}.$$

The higher standard deviation or a variance value is, the more diversified is the group in relation to an analysed feature.

Note

The sample standard deviation is a kind of approximation (estimator) of the population standard deviation. The population standard deviation value is included in a range which contains the sample standard

deviation. This range is called a **confidence interval** for standard deviation.

Coefficient of variation

Coefficient of variation, just like standard deviation, enables you to estimate the homogeneity level of an analysed data collection. It is formulated as:

$$V = \frac{sd}{\bar{x}} 100\%,$$

where *sd* means standard deviation, \bar{x} means arithmetic mean.

This is a unitless value. It enables you to compare a diversity of several different datasets of a one feature. And also, you are able to compare a diversity of several features (expressed in different units). It is assumed, if *V* coefficient does not exceed 10%, features indicate a statistically insignificant diversity.

Standard errors – they are not measures of a measurement dispersion. They measure an accuracy level, you can define the population parameters value, having just the sample estimators. Standard error of the mean is defined by:

$$SEM = \text{standard error of the mean} = \frac{sd}{\sqrt{n}}.$$

Note

On the basis of a sample estimator you can calculate a **confidence interval** for a population parameter.

7.4 ANOTHER DISTRIBUTION CHARACTERISTICS

Skewness or asymmetry coefficient in other words

This measure tells us how **data distribution** differs from symmetrical distribution. The closer the value of skewness is to zero, the more symmetrically around the mean the data are spread. Usually the value of this coefficient is included in a range [-1, 1], but in the case of a very big asymmetry, it may occur outside the above-mentioned range. A positive skew value indicates that the right skew occurs (the tail on the right side is longer), whereas the negative skew indicates that the left skew occurs (the tail on the left side is longer). Skewness is defined by:

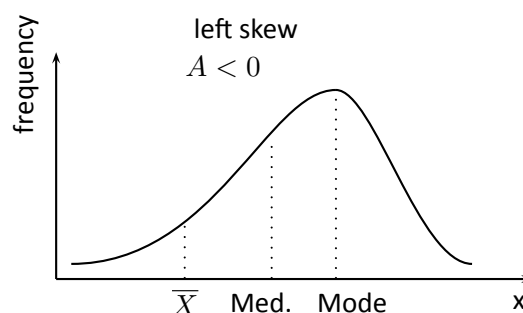
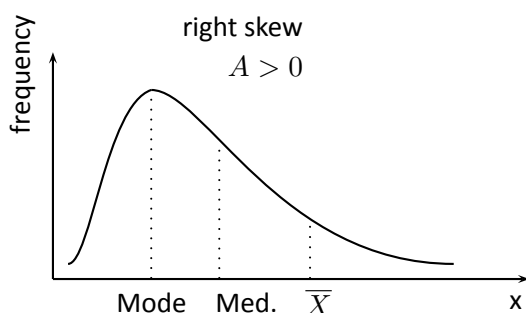
$$A = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{sd} \right)^3,$$

where:

x_i – the following values of a variable,

\bar{x} , *sd* – adequately - arithmetic mean and standard deviation x_i ,

n – sample size.



Kurtosis or coefficient of concentration

This measure tells us how much the spread of data around the mean is similar to the spread of data in **normal distribution**. The greater than zero the value of kurtosis is, the more narrow the tested distribution than normal one is. And inversely, the lower than zero the value of kurtosis is, the flatter the tested distribution than the normal one is. Kurtosis is defined by:

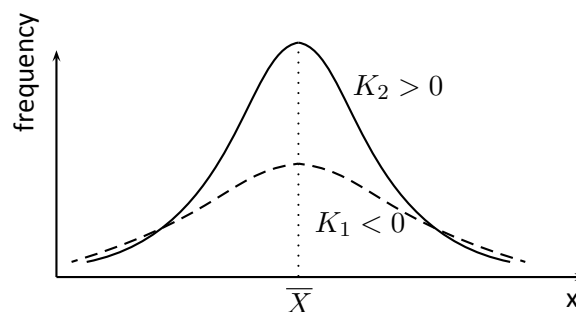
$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{sd} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)},$$

where:

x_i — the following values of a variable,

\bar{x}, sd — adequately - arithmetic mean and standard deviation of x_i ,

n — sample size.

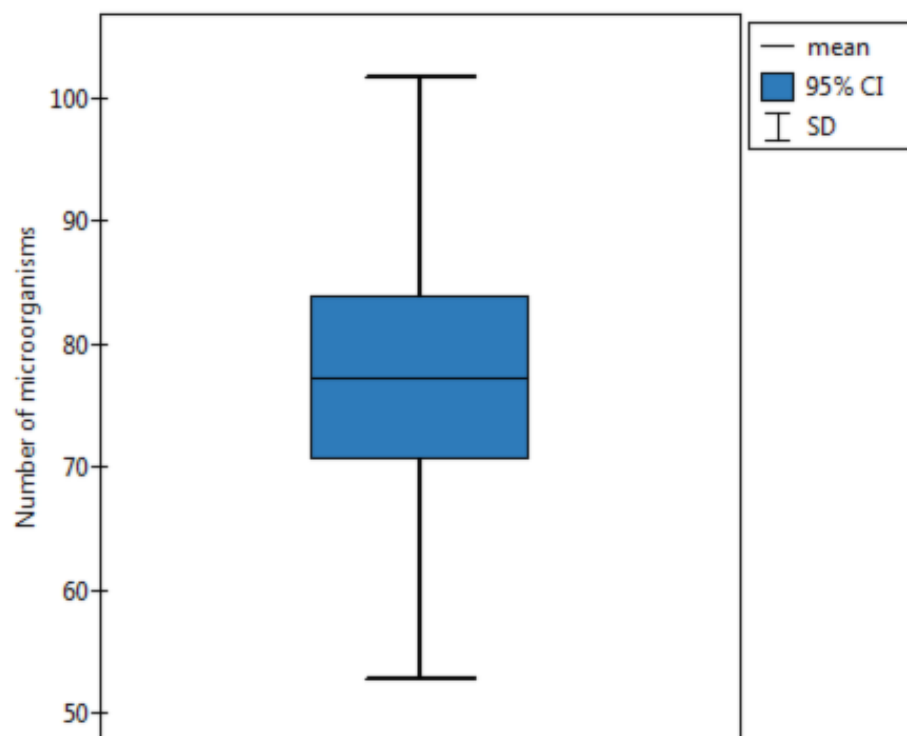


EXAMPLE 7.1. (fertilisers.pqs file)

In an experiment related to a soil fertilising the with various sorts of microbiological specimens and fertilisers it was calculated how many microorganisms occur in a 1 gramme of dry mass of soil. Now we would like to calculate descriptive statistics of the amount of actinomycetes for the sample fertilised with nitrogen. Additionally, we want the data to be illustrated in the Box-Whiskers plot. In a datasheet, we select only the 54 first rows, which are relevant to the assumptions of the analysis (there are actinomycetes fertilised with nitrogen). Then we open Descriptive statistics window in Statistics menu→Descriptive statistics.

In the window of descriptive statistics options, select a variable to analyse: the number of microorganisms, and then all the procedures you want to follow (for example arithmetic mean altogether with the confidence interval, median, standard deviation altogether with the confidence interval, and an information about the skewness and kurtosis of distribution altogether with errors). At the top of the window you should see the following message: **Data reduced by the selected area**. To add a graph to the report, we select Add graph option and chose the Box-Whiskers plot type . Confirm your choice by clicking OK and you get the result in a report:

Descriptive statistics	
Analysis time	0.02sec.
Analysed variables	Number of microorganisms
significance level	0.05
Group size	54
Arithmetic mean	77.240741
Median	78.5
Standard deviation	24.425424
95% Confidence interval for the std. dev.	20.532603
	30.153531
95% Confidence interval for the mean	70.573883
	83.907598
Skewness	-0.226875
Std. err. of the skewness	0.324556
Kurtosis	0.343163
Std. err. of the kurtosis	0.638893

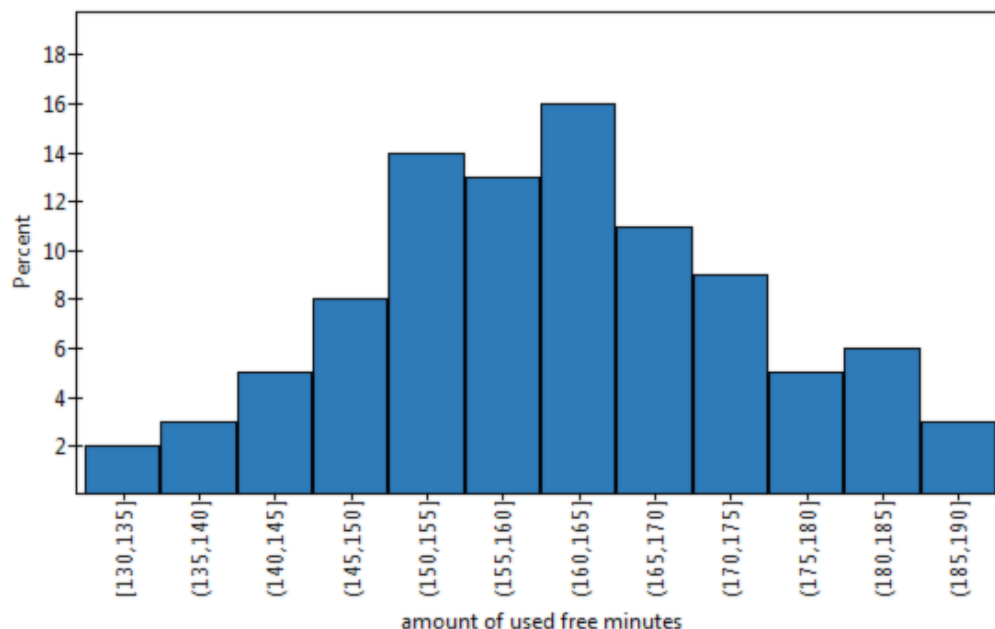


8 PROBABILITY DISTRIBUTIONS

A real data distribution from a sample - **empirical data distribution** may be carried out in a mean of a **frequency tables** (by selecting Statistic menu→Frequency tables). For example, a distribution of the amount of used free minutes by subscribers of some mobile network operator (*example (6.1), distribution.pqs file*) performs the following table:

Frequency tables					
Analysis time		0.02sec.			
Variable: amount of used free minutes	Frequency	Cumulative frequency	Percent	Cumulative percent	
[130,135]	5	5	2.5%	2.5%	
(135,140]	7	12	3.5%	6%	
(140,145]	11	23	5.5%	11.5%	
(145,150]	17	40	8.5%	20%	
(150,155]	29	69	14.5%	34.5%	
(155,160]	27	96	13.5%	48%	
(160,165]	32	128	16%	64%	
(165,170]	23	151	11.5%	75.5%	
(170,175]	19	170	9.5%	85%	
(175,180]	11	181	5.5%	90.5%	
(180,185]	13	194	6.5%	97%	
(185,190]	6	200	3%	100%	

A graphical presentation of results included in a table is usually done using a histogram or a bar plot.



Such graph can be created by selecting Add graph option in the Frequency tables window.

Theoretical data distribution which is also called a **probability distribution** is usually presented graphically by means of a line graph. Such line is described by a function (mathematical model) and it is called

a **density function**. You can replace the empirical distribution with the adequate theoretical distribution.

Note

To replace an empirical distribution with the adequate theoretical distribution it is not enough to draw conclusions upon similarity of their shapes intuitively. To check it, you should use specially created [compatibility tests](#).

The kind of probability distribution which is used the most often is a [normal distribution](#) (Gaussian distribution). Such distribution with a mean of 161.15 and a standard deviation 13.03 is presented by the data relating to the amount of used free minutes (*example (6.1), distribution.pqs file*).

8.1 CONTINUOUS PROBABILITY DISTRIBUTIONS

- **Normal distribution** which is also called the Gaussian distribution or a bell curve, is one of the most important distribution in statistics. It has very interesting mathematical features and occurs very often in nature. It is usually designated with $N(\mu, \sigma)$.

A density function is defined by:

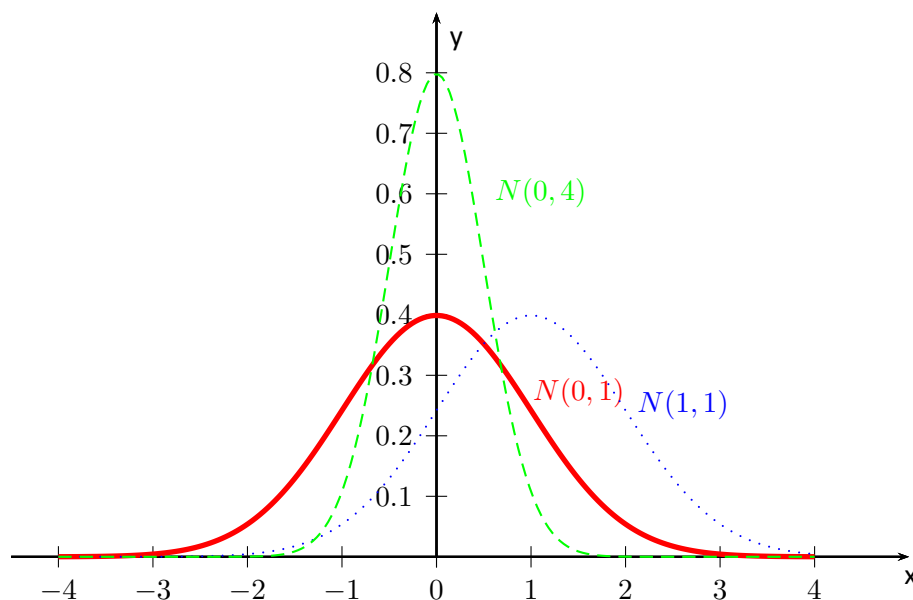
$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right),$$

where:

$$-\infty < x < +\infty,$$

μ – an expected value of population (its measure is mean),

σ – standard deviation.



Normal distribution is a symmetrical distribution for a perpendicular line to axis of abscissae going through the points designating the mean, mode and median.

Normal distribution with a mean of $\mu = 0$ and $\sigma = 1$ ($N(0, 1)$), is so called a **standardised normal distribution**.

- **t-Student distribution** – the shape of t-Student distribution is similar to standardised normal distribution, but its tails are longer. The higher the number of degrees of freedom (df), the more similar the shape of t-Student distribution to normal distribution.

A density function is defined by:

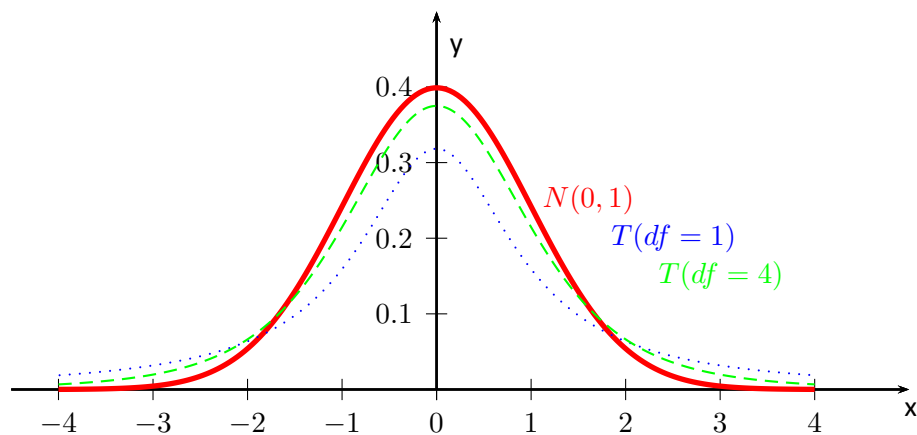
$$f(x, df) = \frac{\Gamma(\frac{df+1}{2})}{\Gamma(\frac{df}{2})\sqrt{df\pi}} \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}},$$

where:

$$-\infty < x < +\infty,$$

df – degrees of freedom (sample size is decreased by the number of limitations in given calculations),

Γ is a Gamma function.



- **Chi-square (χ^2) distribution**, this is a right-skewed distribution with a shape depending on the number of degrees of freedom df . The higher the number of degrees of freedom, the more similar the shape of χ^2 distribution to the normal distribution.

Density function is defined by:

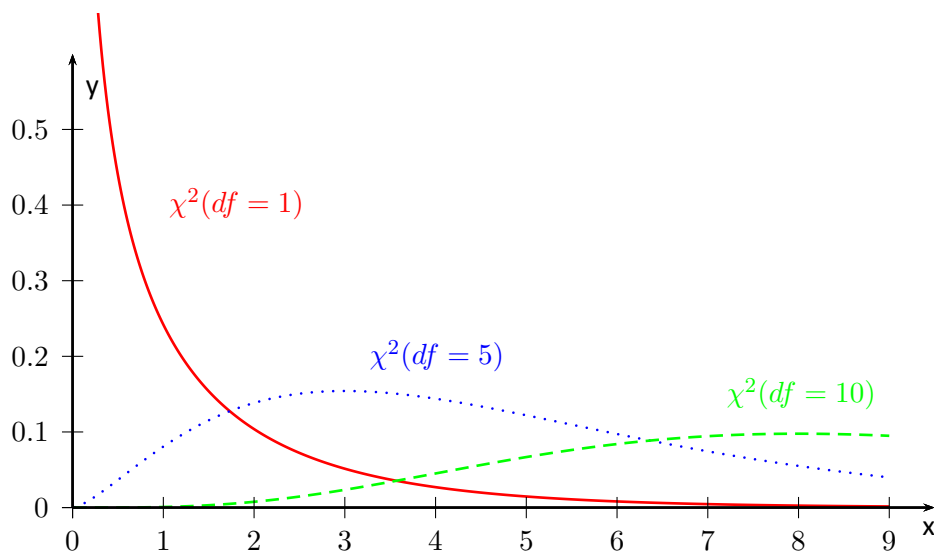
$$f(x, df) = \frac{1}{2^{\frac{df}{2}} \Gamma(\frac{df}{2})} x^{\frac{df}{2}-1} e^{-\frac{x}{2}},$$

where:

$x > 0$,

df – degrees of freedom (sample size is decreased by the number of limitations in given calculations),

Γ is a Gamma function.



- **Fisher-Snedecor distribution**, this is a distribution which has a right tail that is longer and a shape that depends on the number of degrees of freedom df_1 and df_2 .

A density function is defined by:

$$F(x, df_1, df_2) = \frac{\sqrt{\frac{(df_1 x)^{df_1} df_2^{df_2}}{(df_1 x + df_2)^{df_1 + df_2}}}}{x B\left(\frac{df_1}{2}, \frac{df_2}{2}\right)},$$

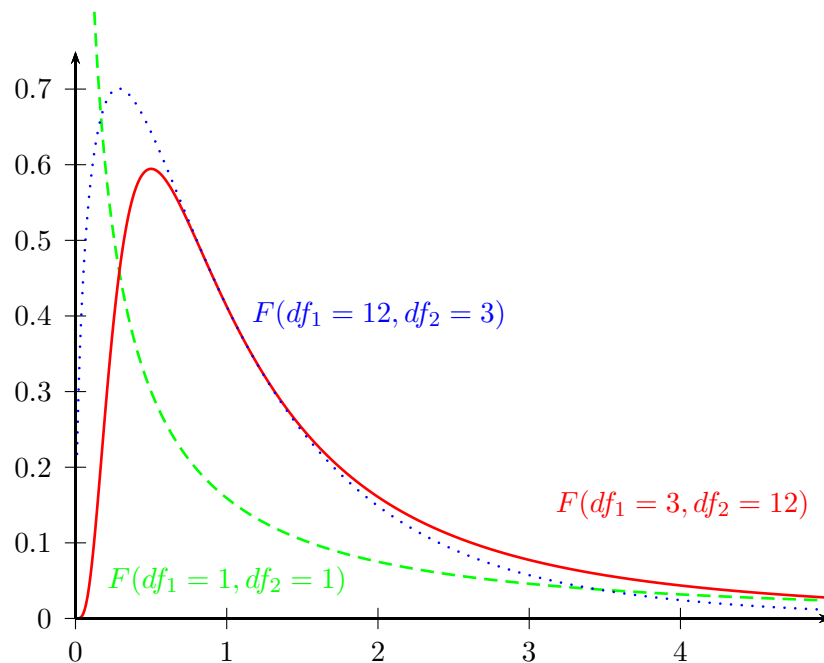
where:

$x > 0$,

df_1, df_2 — degrees of freedom (it is assumed that if X i Y are independent with a χ^2 distribution with adequately df_1 and df_2 degrees of freedom, then $F = \frac{X/df_1}{Y/df_2}$ has a F

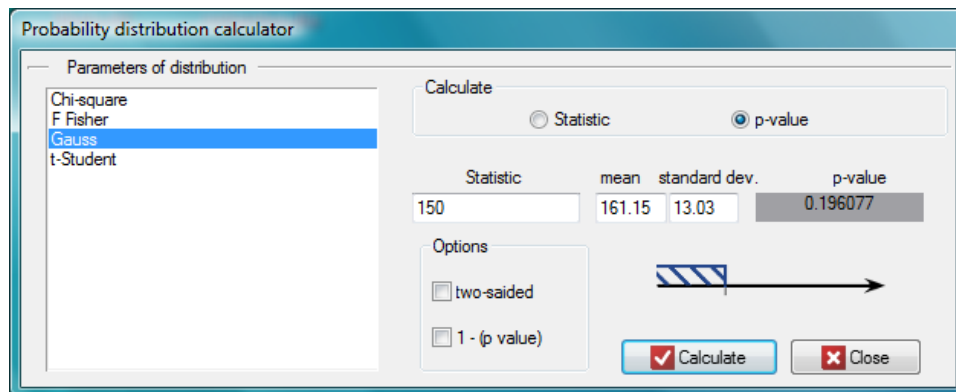
Snedecor distribution $F(df_1, df_2)$),

B is a Beta function.



8.2 PROBABILITY DISTRIBUTION CALCULATOR

The area under a curve (density function) is p **probability** of occurrence of all possible values of an analysed random variable. The whole area under a curve comes to $p = 1$. If you want to analyse just a part of this area, you must put the border value, which is called the **critical value** or Statistic. To do this, you need to open the Probability distribution calculator window. In this window you can calculate not only a value of the area under the curve (p value) of the given distribution on the basis of Statistic, but also Statistic value on the basis of p value. To open the window of Probability distribution calculator, you need to select Probability distribution calculator from the Statistics menu.



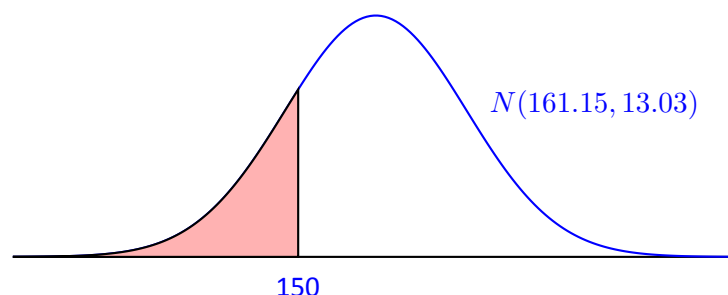
EXAMPLE 8.1. Probability distribution calculator

Some mobile network operator did the research, which was supposed to show the usage of "free minutes" given to his clients on a pay-monthly contract. On the basis of the sample, which consists of 200 of the above-mentioned network clients (where the distribution of used free minutes is of the shape of **normal distribution**) is calculated the mean value $\bar{x} = 161.15min.$ and standard deviation $sd = 13.03min.$ We want to calculate the probability, that the chosen client used:

1. 150 minutes or less,
2. more than 150 minutes,
3. the amount of minutes coming from the range $[\bar{x} - sd, \bar{x} + sd] = [148.12min., 174.18min.]$,
4. the amount of minutes out of the range $\bar{x} \pm sd$.

Open the Probability distribution calculator window, select Gaussian distribution and write the mean $\bar{x} = 161.15min.$ and standard deviation $sd = 13.03min.$ and select the option which indicates, that you are going to calculate the p value.

1. To calculate (using normal distribution (Gauss)) the probability that the client you have chosen used 150 free minutes or less, put the value of 150 in the Statistic field. Confirm all selected settings by clicking Calculate.

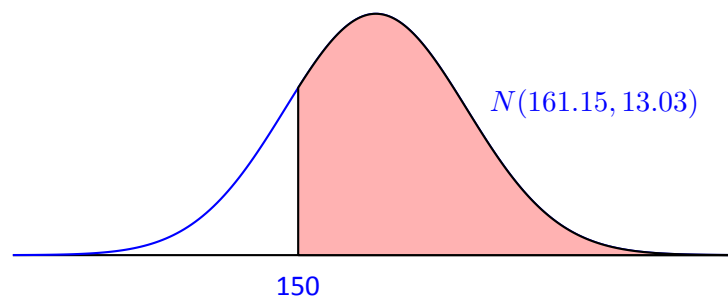


The obtained p value is 0.193961.

Note

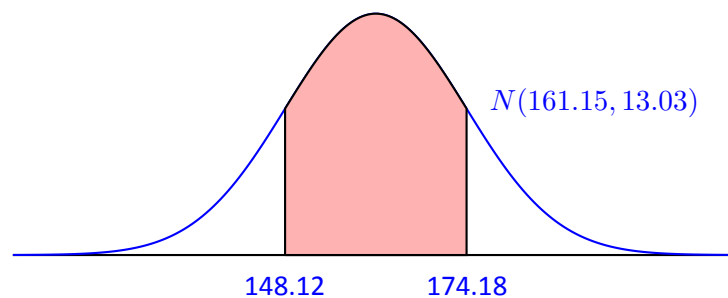
Similar calculations you can carry out on the basis of empirical distribution. The only thing you should do is to calculate a percentage of clients who use 150 minutes or less (*example (6.1)* by using the Frequency tables window. In the analysed sample (which consists of 200 clients) there are 40 clients who use 150 minutes or less. It is 20% of the whole sample, so the probability you are looking for is $p = 0.2$.

- To calculate the probability (using the normal distribution (Gauss)), that the client who you have chosen used more than 150 free minutes, you need to put the value of 150 in the Statistic field and then select the option 1 - (p value). Confirm all the chosen settings by clicking Calculate.



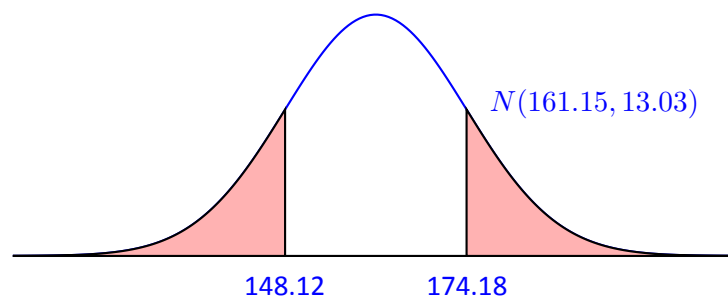
The obtained p value is 0.806039.

- To calculate (using the normal distribution (Gauss)) a probability that the client you have chosen used free minutes which come from the range $[\bar{x} - sd, \bar{x} + sd] = [148.12min., 174.18min.]$ in the Statistic field, put one of the final range values and then select the option two-sided. Confirm all the chosen settings by clicking Calculate.



The obtained p value is 0.682689.

- To calculate (using the normal distribution (Gauss)) a probability, that the client you have chosen used free minutes out of the range $[\bar{x} - sd, \bar{x} + sd] = [148.12min., 174.18min.]$ in the Statistic field put one of the final range values and then select the option: two-sided and 1 - (p value). Confirm all the chosen settings by clicking Calculate.





The obtained p value is 0.317311.

9 HYPOTHESES TESTING

The process of generalisation of the results obtained from the sample for the whole population is divided into 2 basic parts:

- **estimation** — estimating values of the parameters of the population on the basis of the statistical sample,
- **verification of statistical hypotheses** — testing some specific assumptions formulated for the parameters of the general population on the basis of sample results.

9.0.1 POINT AND INTERVAL ESTIMATION

In practice, we usually do not know the **parameters** (characteristics) of the whole population. There is only a sample chosen from the population. **Point estimators** are the characteristics obtained from a random sample. The exactness of the estimator is defined by its **standard error**. The real parameters of population are in the area of the indicated point estimator. For example, the population parameter **arithmetic mean** μ is in the area of the estimator from the sample which is \bar{x} .

If you know the estimators of the sample and their theoretical distributions, you can estimate values of the population parameters with the **confidence level** $(1 - \alpha)$ defined in advance. This process is called **interval estimation**, the interval: **confidence interval**, and α is called a **significance level**.

The most popular significance level comes to 0.05, 0.01 or 0.001.

9.0.2 VERIFICATION OF STATISTICAL HYPOTHESES

To verify a statistical hypotheses, follow several steps:

The 1st step: Make a hypotheses, which can be verified by means of statistical **tests**.

Each statistical test gives you a general form of the null hypothesis \mathcal{H}_0 and the alternative one \mathcal{H}_1 :

\mathcal{H}_0 : there is **no** statistically significant **difference** among **populations**
(means, medians, proportions distributions etc.),

\mathcal{H}_1 : there **is** a statistically significant **difference** among **populations**
(means, medians, proportions, distributions etc.).

Researcher must formulate the hypotheses in the way, that it is compatible with the reality and statistical **test** requirements, for example:

\mathcal{H}_0 : the percentage of women and men running their own businesses
in an analysed population is exactly the same.

If you do not know, which percentage (men or women) in an analysed population might be greater, the alternative hypothesis should be two-sided. It means you should not assume the direction:

\mathcal{H}_1 : the percentage of women and men running their own businesses
in an analysed population is different.

It may happen (but very rarely) that you are sure you know the direction in an alternative hypothesis. In this case you can use one-sided alternative hypothesis.

The 2nd step: Verify which of the hypotheses \mathcal{H}_0 or \mathcal{H}_1 is more probable. Depending on the kind of an analysis and a type of variables you should choose an appropriate statistical [test](#).

Note 1

Note, that choosing a statistical test means mainly choosing an appropriate [measurement scale](#) (interval, ordinal, nominal scale) which is represented by the data you want to analyse. It is also connected with choosing the analysis model (dependent or independent)

Measurements of the given feature are called **dependent (paired)**, when they are made a couple of times for the same objects. When measurements of the given feature are performed on the objects which belong to different groups, these groups are called **independent (unpaired)** measurements.

Some examples of researches in dependent groups:

Examining a body mass of patients before and after a slimming diet, examining reaction on the stimulus within the same group of objects but in two different conditions (for example - at night and during the day), examining the compatibility of evaluating of credit capacity calculated by two different banks but for the same group of clients etc.

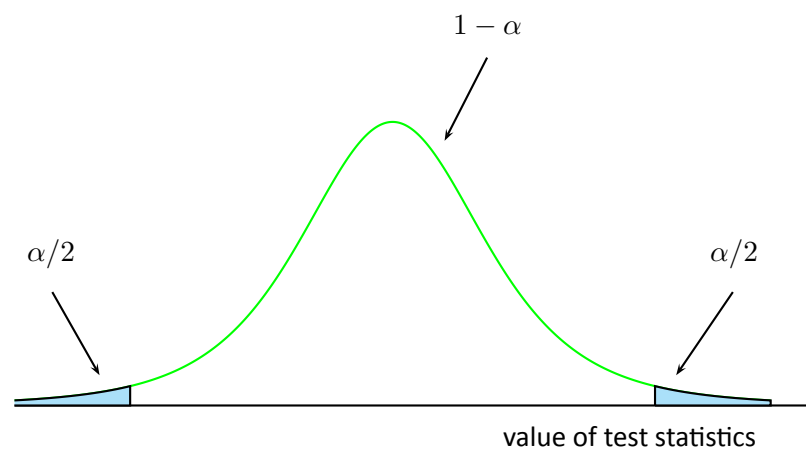
Some examples of researches in independent groups:

Examining a body mass in a group of healthy patients and ill ones, testing effectiveness of fertilising several different kinds of fertilisers, testing gross domestic product (GDP) sizes for the several countries etc.

Note 2

A graph which is included in the [Wizard](#) window makes the choice of an appropriate statistical test easier.

Test statistic of the selected test calculated according to its formula is connected with the adequate theoretical distribution.



The application calculates a value of [test statistics](#) and also a [p value](#) for this statistics (a part of the area under a curve which is adequate to the value of the test statistics). The [p value](#) enables

you to choose a more probable hypothesis (null or alternative). But you always need to assume if a null hypothesis is the right one, and all the proofs gathered as a data are supposed to supply you with the enough number of counterarguments to the hypothesis:

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

There is usually chosen **significance level** $\alpha = 0.05$, accepting that for 5 % of the situations we will reject a null hypothesis if there is the right one. In specific cases you can choose other significance level for example 0.01 or 0.001.

Note

Note, that a statistical test may not be compatible with the reality in two cases:

		reality	
		$\mathcal{H}_0 : \text{true}$	$\mathcal{H}_0 : \text{false}$
test result	$\mathcal{H}_0 : \text{true}$	OK	β
	$\mathcal{H}_0 : \text{false}$	α	OK

We may make two kinds of mistakes:

$\alpha =$ **1st type of error** (probability of rejecting hypothesis \mathcal{H}_0 , when it is the right one),

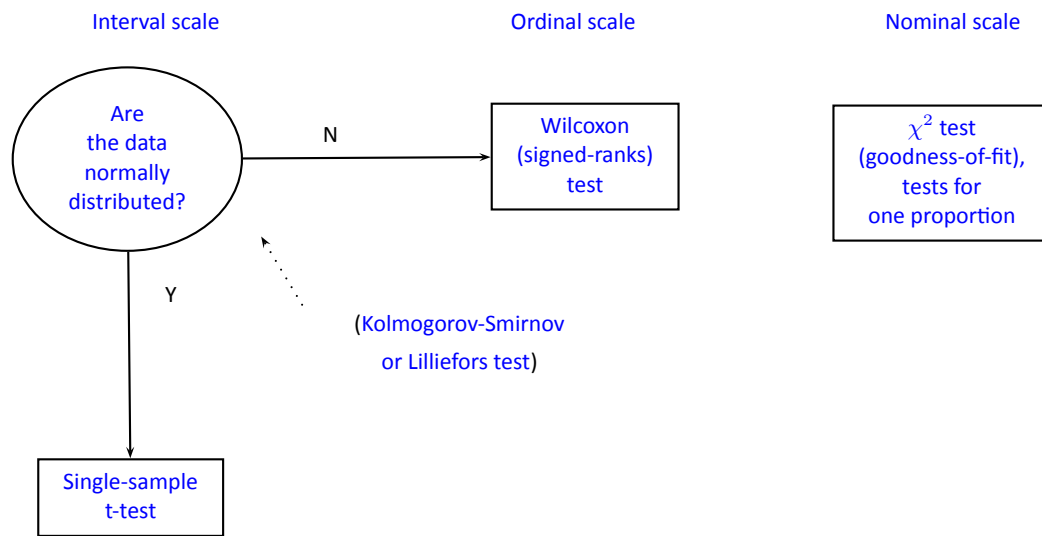
$\beta =$ 2nd type of error (probability of accepting hypothesis \mathcal{H}_0 , when it is the wrong one).

Power of the test is $1 - \beta$.

Values α and β are connected with each other. The approved practice is to assume the significance level in advance α and minimalization β by decreasing a sample size.

The 3rd step: Description of results of hypotheses verification.

10 COMPARISON - 1 GROUP



10.1 PARAMETRIC TESTS

10.1.1 The t -test for a single sample

The single-sample t test is used to verify the hypothesis, that an analysed sample with the mean (\bar{x}) comes from a population, where mean (μ) is a given value.

Basic assumptions:

- measurement on an [interval scale](#),
- [normality of distribution](#) of an analysed feature.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \mu = \mu_0, \\ \mathcal{H}_1 : & \mu \neq \mu_0,\end{aligned}$$

where:

μ – mean of an analysed feature of the population represented by the sample,
 μ_0 – a given value.

The test statistic is defined by:

$$t = \frac{\bar{x} - \mu_0}{sd} \sqrt{n},$$

where:

sd – standard deviation from the sample,
 n – sample size.

The test statistic has the [t-Student distribution](#) with $n - 1$ degrees of freedom.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha & \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

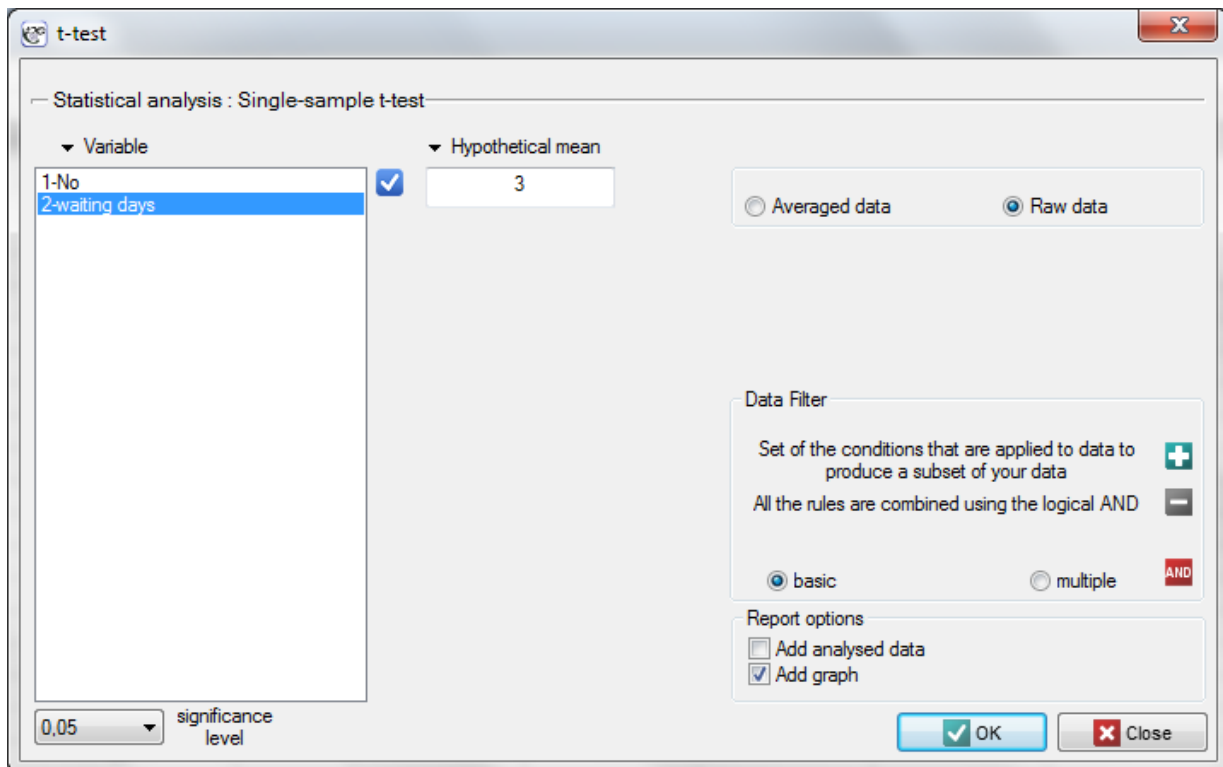
Note

Note, that: If the sample is large and you know a standard deviation of the population, then you can calculate a test statistic using the formula:

$$t = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}.$$

The statistic calculated this way has the [normal distribution](#). If $x \rightarrow \infty$ t -Student distribution converges to the normal distribution $N(0, 1)$. In practice, it is assumed, that with $n > 30$ the t -Student distribution may be approximated with the normal distribution.

The settings window with the Single-sample t -test can be opened in Statistics menu→Parametric tests→t-test or in [Wizard](#).



Note

Calculations can be based on **raw data** or data that are averaged like: arithmetic mean, standard deviation and sample size.

EXAMPLE 10.1. (courier.pqs file)

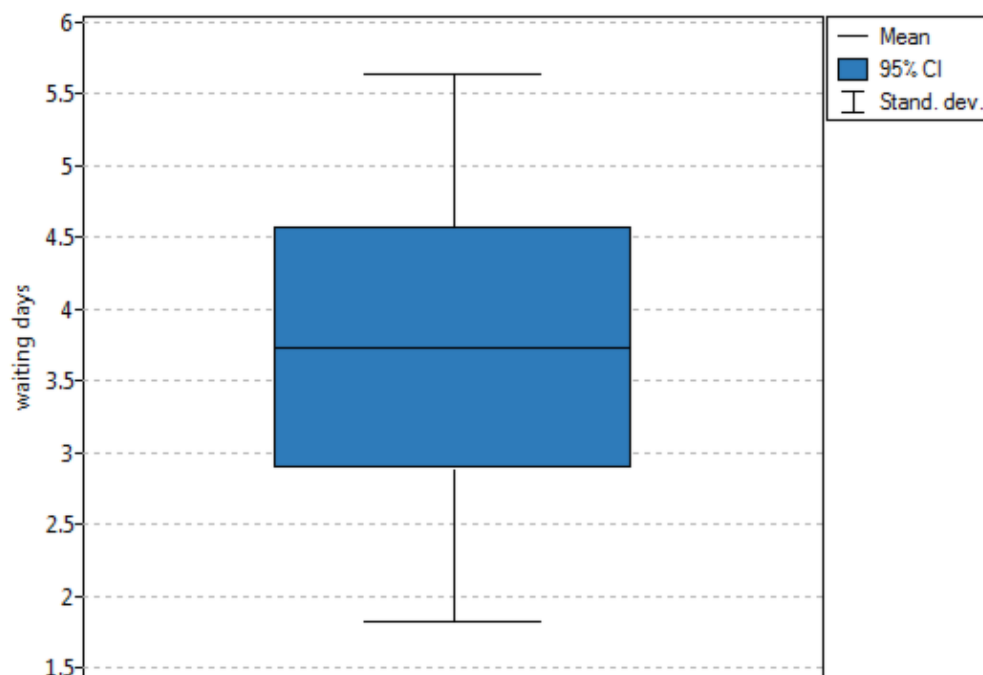
You want to check if the time of awaiting for a delivery by some courier company is 3 days on the average ($\mu_0 = 3$). In order to calculate it, there are 22 persons chosen by chance from all clients of the company as a sample. After that, there are written information about the number of days passed since the delivery was sent till it is delivered. There are following values: (1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7).

The number of awaiting days for the delivery in the analysed population fulfills the assumption of **normality** of distribution.

Hypotheses:

- \mathcal{H}_0 : mean of the number of awaiting days for the delivery, which is supposed to be delivered by the above-mentioned courier company is 3,
- \mathcal{H}_1 : mean of the number of awaiting days for the delivery, which is supposed to be delivered by the above-mentioned courier company is different from 3.

Single-sample t-test	
Analysis time	0.03sec.
Analysed variables	waiting days
Significance level	0.05
Group size	22
Hypothetical mean	3
Group mean	3.727273
Std. err. of the group mean	0.406558
Group standard deviation	1.906925
-95% CI for the group mean	2.881789
+95% CI for the group mean	4.572756
Difference of the means	0.727273
-95% CI for the difference	-0.118211
+95% CI for the difference	1.572756
t-statistic	1.788854
Degrees of freedom	21
p-value	0.088074



Comparing the p value = 0.088074 of the t -test with the significance level $\alpha = 0.05$ we draw the conclusion, that there is no reason to reject the null hypothesis which informs that the average time of awaiting for the delivery, which is supposed to be delivered by the analysed courier company is 3. For the tested sample, the mean is $\bar{x} = 3.727$ and the standard deviation is $sd = 1.907$.

10.2 NONPARAMETRIC TESTS

Ranks - there are the following numbers (usually natural ones) ascribed to the values of ordered measurements of the analysed variable. They are usually used in such nonparametric tests, which are based only upon the order of elements in the sample. Replacing a variable with the grades calculated on the basis of this variable is called **ranking**.

All reoccurring values have its own ascribed rank, which is an arithmetic mean calculated from the following natural numbers proposed to these values. These kinds of ranks are called **ties**.

For example, to the variable of the following values: 8.6, 5.3, 8.6, 7.1, 9.3, 7.2, 7.3, 7.4, 7.3, 5.2, 7, 9.9, 8.6, 5.7 the following ranks are ascribed:

sorted values of variable	ranks
5.2	1
5.3	2
5.7	3
7	4
7.1	5
7.2	6
7.3	7.5
7.3	7.5
7.4	9
8.6	11
8.6	11
8.6	11
9.3	13
9.9	14

But, to the variable with the values of 7.3 is ascribed the tie calculated from the numbers: 7 and 8, and to the variable with the values of 8.6 the tie is calculated from the numbers: 10, 11, 12.

10.2.1 The Kolmogorov-Smirnov test and the Lilliefors test

The Kolmogorov-Smirnov goodness-of-fit test for a single sample, Kolmogorov (1933)[45], is used to verify the hypothesis about the insignificance difference of an analysed variable distribution (*empirical distribution*) from the normal distribution (*theoretical distribution*). We use it in the situation when a mean value (μ) and standard deviation (σ) of the population from which the sample is taken, are known. When these parameters of the population are not known but are estimated and based on the sample, the Kolmogorov test becomes pretty conservative (using this test it is much harder to reject null hypothesis). In such situation you should use the Lilliefors test, Lilliefors (1967, 1969, 1973)[51][52][53]. This is the Kolmogorov-Smirnov test correction when a mean value(μ) and standard deviation (σ) of the population from which the sample is taken, are not known.

Basic assumptions:

- measurement on an [interval scale](#).

Hypotheses:

- \mathcal{H}_0 : distribution of an analysed feature in the population is the normal distribution,
- \mathcal{H}_1 : distribution of an analysed feature in the population is different from the normal one.

Based on the data from the sample gathered in a cumulated frequency distribution and the adequate values of the area under a theoretical curve of the normal distribution, you can calculate a value of the test statistic D :

$$D = \sup_x |F_n(x) - F(x)|,$$

where:

$F_n(x)$ – empirical cumulative distribution function of the normal distribution, calculated in particular points of distribution, for sample of n -elements ,

$F(x)$ – theoretical cumulative distribution function of the normal distribution.

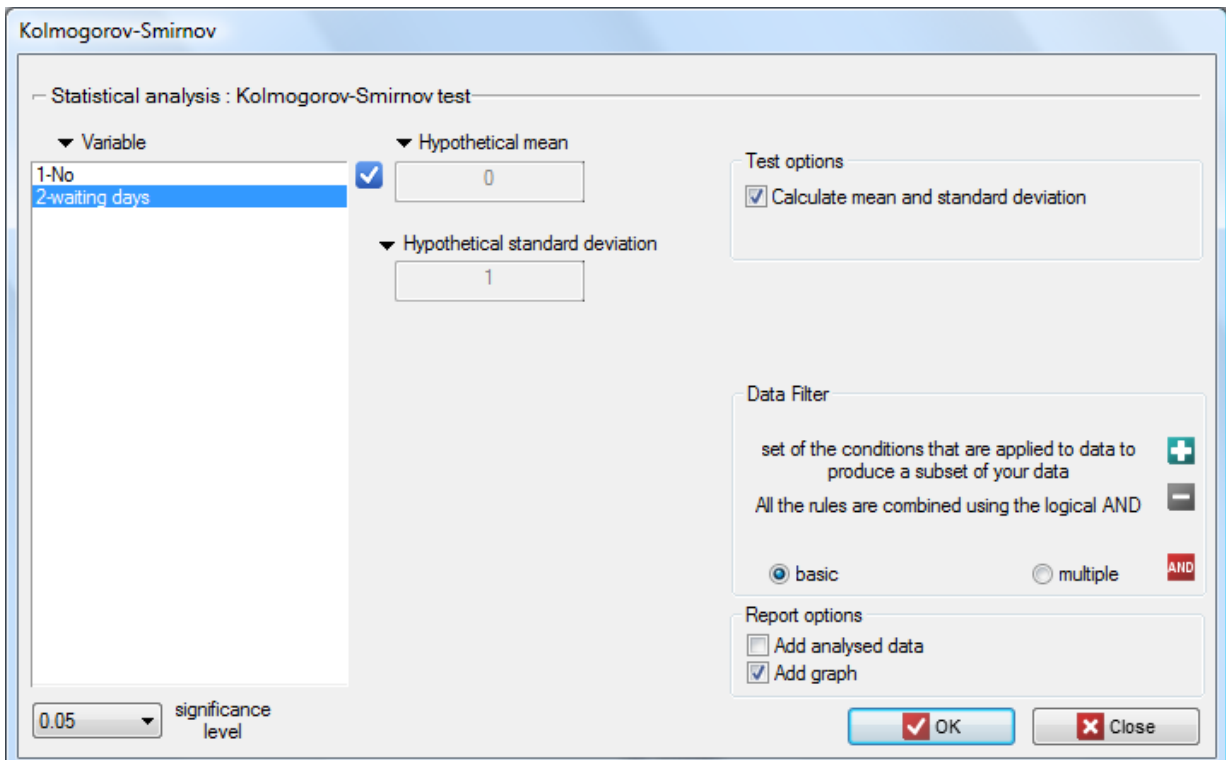
This statistic has the Kolmogorov-Smirnov distribution (if you know the arithmetic mean and the standard deviation of the population) or the Lilliefors distribution (when the arithmetic mean and the standard deviation you want to estimate from the sample).

The p value, designated on the basis of the test statistic, is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 accept \mathcal{H}_1 ,

if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

The settings window with the Lilliefors test or Kolmogorov-Smirnov test can be opened in Statistics menu → NonParametric tests (ordered categories) or in [Wizard](#).



The screenshot shows the 'Kolmogorov-Smirnov' settings window. The 'Statistical analysis' section is set to 'Kolmogorov-Smirnov test'. Under 'Variable', '2-waiting days' is selected. The 'Hypothetical mean' is set to 0 and the 'Hypothetical standard deviation' is set to 1. The 'Test options' section has 'Calculate mean and standard deviation' checked. The 'Data Filter' section shows 'set of the conditions that are applied to data to produce a subset of your data' with a '+' icon, and 'All the rules are combined using the logical AND' with a '-' icon. The 'Report options' section has 'Add analysed data' unchecked and 'Add graph' checked. The 'significance level' is set to 0.05. The 'OK' button is highlighted with a red checkmark, and the 'Close' button has a red X icon.

Lilliefors

Statistical analysis : Lilliefors test

Variable

1-No ☒

2-waiting days

0.05 significance level

Data Filter

set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

☒ basic ☐ multiple

Report options

☐ Add analysed data

☒ Add graph

OK Close

EXAMPLE 10.1 continuation (courier.pqs file)

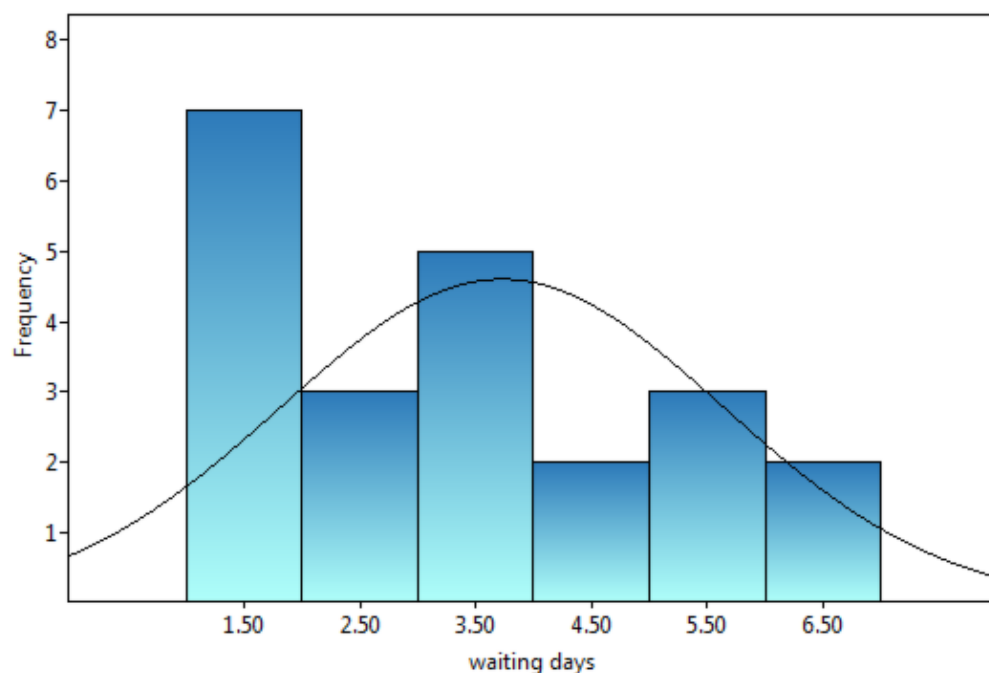
Hypotheses:

- \mathcal{H}_0 : distribution of the number of awaiting days for the delivery, which is supposed to be delivered by the analysed courier company is the normal distribution,
- \mathcal{H}_1 : distribution of the number of awaiting days for the delivery, which is supposed to be delivered by the analysed courier company is different from the normal distribution.

The mean value and the standard deviation of the time of awaiting for the delivery for all the clients is not known, so it must be estimated from the sample. There are following values for this sample: $\bar{x} = 3.73$, $SD = 1.91$.

Kolmogorov-Smirnov test (goodness-of-fit)	
Analysis time	0.02sec.
Analysed variables	waiting days
significance level	0.05
Group size	22
Group mean	3.727273
Group standard deviation	1.906925
D statistic	0.135658
Degrees of freedom	22
p-value	0.763881

Lilliefors test (goodness-of-fit)	
Analysis time	0.02sec.
Analysed variables	waiting days
significance level	0.05
Group size	22
Group mean	3.727273
Group standard deviation	1.906925
D statistic	0.135658
Degrees of freedom	22
p-value	0.364381



The value of the Kolmogorov-Smirnov and the Lilliefors test statistic is exactly the same and amounts to 0.1357, but the p value = 0.763881 for the Kolmogorov-Smirnov test and the p value = 0.364381 for Lilliefors test. Both tests indicate, that using the significance level $\alpha=0.05$ you are not allowed to reject the null hypothesis which informs, that the analysed data performs the normal distribution.

10.2.2 The Wilcoxon test (signed-ranks)

The Wilcoxon signed-ranks test is also known as the Wilcoxon single sample test, Wilcoxon (1945, 1949)[83]. This test is used to verify the hypothesis, that the analysed sample comes from the population, where median (θ) is a given value.

Basic assumptions:

- measurement on an [ordinal scale](#) or on an [interval scale](#).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : \theta &= \theta_0, \\ \mathcal{H}_1 : \theta &\neq \theta_0.\end{aligned}$$

where:

θ – median of an analysed feature of the population represented by the sample,

θ_0 – a given value.

Now you should calculate the value of the test statistics Z (T – for the small sample size), and based on this p value.

The p value, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

Note

Depending on the size of the sample, the test statistic takes a different form:

- for a small sample size

$$T = \min \left(\sum R_-, \sum R_+ \right),$$

where:

$\sum R_+$ and $\sum R_-$ are adequately: a sum of positive and negative **rank**s.

This statistic has the Wilcoxon distribution

- for a large sample size

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}},$$

where:

n - the number of ranked signs (the number of ranks),

t - the number of cases being included in the interlinked rank.

The test statistic formula Z includes the correction for **ties**. This correction should be used when ties occur (when there are no ties, the correction is not calculated, because $(\sum t^3 - \sum t) / 48 = 0$).

Z statistic asymptotically (for a large sample size) has the **normal distribution**.

Continuity correction of the Wilcoxon test (Marascuilo and McSweeney (1977)[60])

A continuity correction is used to enable the test statistic to take in all values of real numbers, according to the assumption of the normal distribution. Test statistic with a continuity correction is defined by:

$$Z = \frac{\left| T - \frac{n(n+1)}{4} \right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}}.$$

The settings window with the Wilcoxon test (signed-ranks) can be opened in Statistics menu → Non-Parametric tests (ordered categories) → Wilcoxon (signed-ranks) or in **Wizard**.

Wilcoxon (signed-ranks)

Statistical analysis : Wilcoxon test (signed-ranks)

Variable: 1-No, 2-waiting days

Median: ☒ 3

Test options: ☒ Continuity correction

Data Filter: set of the conditions that are applied to data to produce a subset of your data. All the rules are combined using the logical AND. ☒ basic, ☐ multiple, AND

Report options: ☐ Add analysed data, ☒ Add graph

0.05 significance level

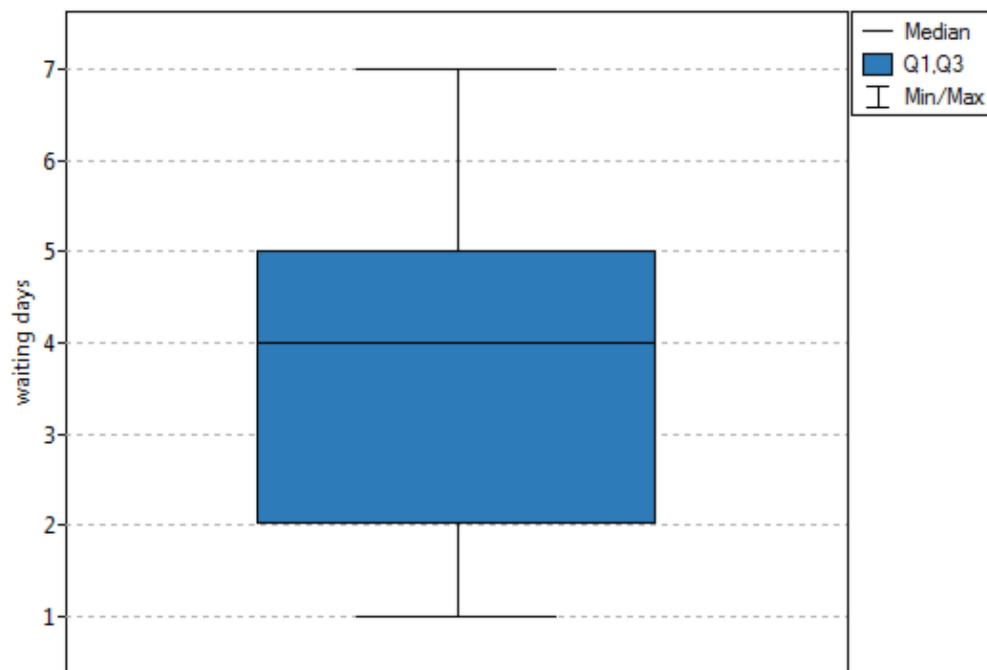
Example 10.1 cont. (*courier.pqs file*)

Hypotheses:

\mathcal{H}_0 : median of the number of awaiting days for the delivery, which is supposed to be delivered by the analysed courier company is 3

\mathcal{H}_1 : median of the number of awaiting days for the delivery, which is supposed to be delivered by the analysed courier company is different from 3

Wilcoxon test (signed-ranks)	
Analysis time	0.03sec.
Analysed variables	waiting days
Significance level	0.05
Continuity correction	Yes
Group size	22
Count of omitted values (equal median)	3
Group median	4
Hypothetical median	3
Sum of negative ranks	134
Sum of positive ranks	56
t statistic	56
p-value (exact)	0.123212
Z statistic (adjusted for ties)	1.572575
p-value (asymptotic)	0.115817



Comparing the p value = 0.123212 of Wilcoxon test based on T statistic with the significance level $\alpha = 0.05$ we draw the conclusion, that there is no reason to reject the null hypothesis informing us, that usually the number of awaiting days for the delivery which is supposed to be delivered by the analysed courier company is 3. Exactly the same decision you would make basing on the p value = 0.111161 or p value = 0.115817 of Wilcoxon test based upon Z statistic or Z with correction for continuity.

10.2.3 The Chi-square goodness-of-fit test

The χ^2 test (goodness-of-fit) is also called the one sample χ^2 test and is used to test the compatibility of values observed for r ($r \geq 2$) categories X_1, X_2, \dots, X_r of one feature X with hypothetical expected values for this feature. The values of all n measurements should be gathered in a form of a table consisted of r rows (categories: X_1, X_2, \dots, X_r). For each category X_i there is written the frequency of its occurrence O_i , and its expected frequency E_i or the probability of its occurrence p_i . The expected frequency is designated as a product of $E_i = np_i$. The built table can take one of the following forms:

X_i categories	O_i	E_i	X_i categories	O_i	p_i
X_1	O_1	E_1	X_1	O_1	p_1
X_2	O_2	E_2	X_2	O_2	p_2
...
X_r	O_r	E_r	X_r	O_r	p_r

Basic assumptions:

- measurement on a **nominal scale** (alternatively: an **ordinal scale** or an **interval scale**),
- large expected frequencies (according to the Cochran interpretation (1952)[20] none of these expected frequencies can be < 1 and no more than 20% of the expected frequencies can be < 5),
- observed frequencies total should be exactly the same as an expected frequencies total, and the total of all p_i probabilities should come to 1.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 &: O_i = E_i \text{ for all categories,} \\ \mathcal{H}_1 &: O_i \neq E_i \text{ for at least one category.}\end{aligned}$$

Test statistic is defined by:

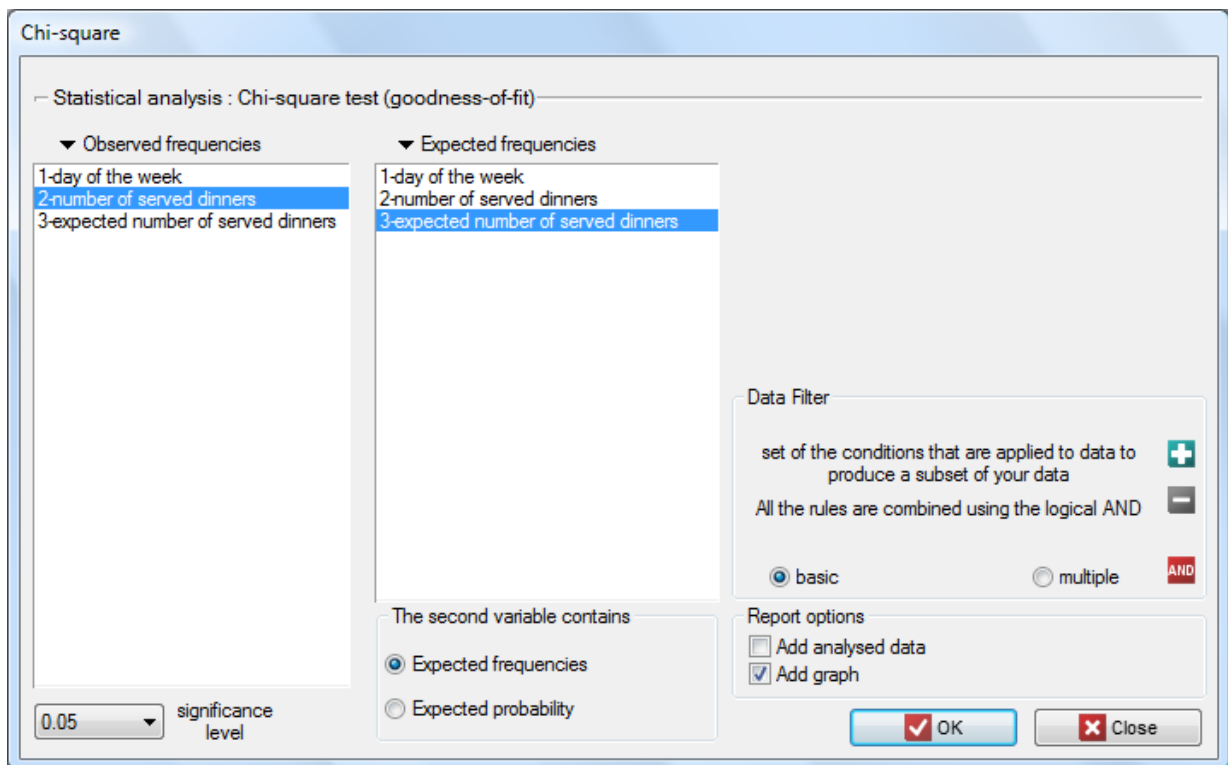
$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with the number of degrees of freedom calculated using the formula: $df = (r - 1)$.

The p value, designated on the basis of the test statistic, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The settings window with the Chi-square test (goodness-of-fit) can be opened in Statistics menu → NonParametric tests (unordered categories) → Chi-square or in Wizard.



EXAMPLE 10.2. (dinners.pqs file)

We would like to get to know if the number of dinners served in some school canteen within a given frame of time (from Monday to Friday) is statistically the same. To do this, there was taken a one-week-sample and written the number of served dinners in the particular days: Monday - 33, Tuesday - 29, Wednesday - 32, Thursday - 36, Friday - 20.

As a result there were 150 dinners served in this canteen within a week (5 days).

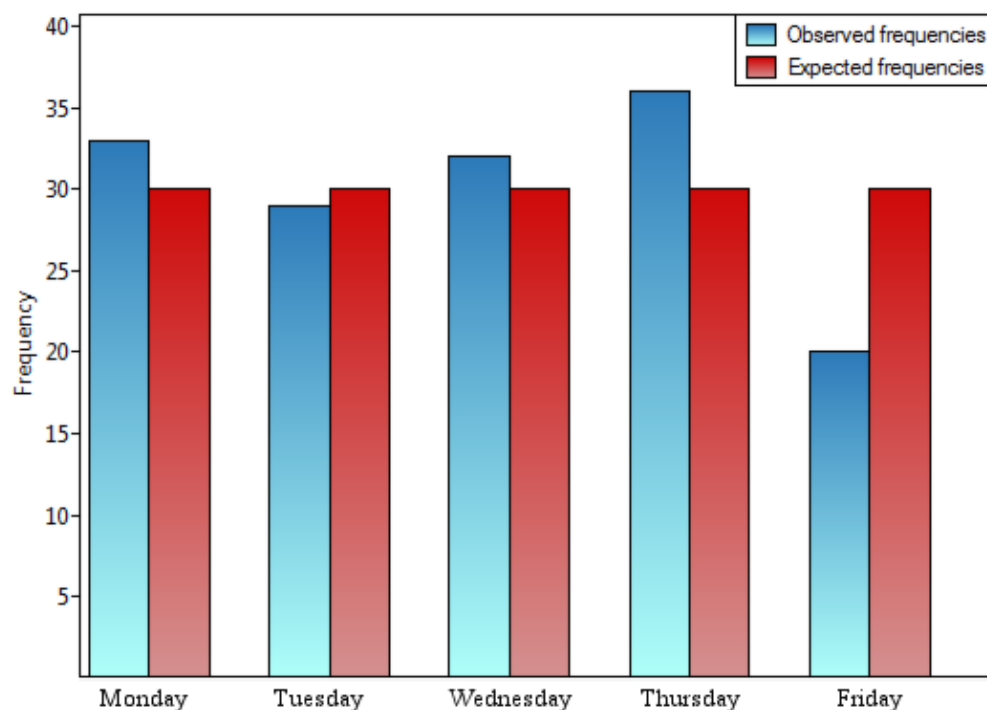
We assume that the probability of serving dinner each day is exactly the same, so it comes to $\frac{1}{5}$. The expected frequencies of served dinners for each day of the week (out of 5) comes to $E_i = 150 \cdot \frac{1}{5} = 30$.

day of the week	number of served dinners	expected number of served dinners
Monday	33	30
Tuesday	29	30
Wednesday	32	30
Thursday	36	30
Friday	20	30

Hypotheses:

- \mathcal{H}_0 : the number of served dinners in the analysed school canteen within given days (of the week) is consistent with the expected number of given out dinners these days,
- \mathcal{H}_1 : the number of served out dinners in the analysed school canteen within a given week is not consistent with the expected number of dinners given out these days.

Chi-square test (goodness-of-fit)	
Analysis time	0.02sec.
Analysed variables	number of served dinners, expected number of se
Significance level	0.05
Size	150
Chi-square statistic	5
Degrees of freedom	4
p-value	0.287297



The p value from the χ^2 distribution with 4 degrees of freedom comes to 0.287297. So using the significance level $\alpha = 0.05$ you can estimate that there is no reason to reject the null hypothesis that informs about the compatibility of the number of served dinners with the expected number of dinners served within the particular days.

Note!

If you want to make more comparisons within the framework of a one research, it is possible to use the **Bonferroni correction**[1]. The correction is used to limit the size of I type error, if we compare the observed frequencies and the expected ones between particular days, for example:

Friday \iff Monday,

Friday \iff Tuesday,

Friday \iff Wednesday,

Friday \iff Thursday,

Provided that, the comparisons are made independently. The significance level $\alpha = 0.05$ for each comparison must be calculated according to this correction using the following formula: $\alpha = \frac{0.05}{r}$, where r is the number of executed comparisons. The significance level for each comparison according to the Bonferroni correction (in this example) is $\alpha = \frac{0.05}{4} = 0.0125$.

However, it is necessary to remember that if you reduce α for each comparison, the **power of the test** is increased.

10.2.4 Tests for proportion

You should use tests for proportion if there are two possible results to obtain (one of them is a distinguished result with the size of m) and you know how often these results occur in the sample (we know a p proportion). Depending on a sample size n you can choose the **Z test for a one proportion** – for large samples and the exact **binominal test for a one proportion** – for small sample sizes. These tests are used to verify the hypothesis that the proportion in the population, from which the sample is taken, is a given value.

Basic assumptions:

- measurement on a **nominal scale** (alternatively: an **ordinal scale** or an **interval scale**).

The additional condition for the Z test for proportion

- large frequencies (according to Marascuilo and McSweeney interpretation (1977)[60] each of these values: $np > 5$ and $n(1 - p) > 5$).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & p = p_0, \\ \mathcal{H}_1 : & p \neq p_0,\end{aligned}$$

where:

p – probability (distinguished proportion) in the population,

p_0 – expected probability (expected proportion).

The Z test for one proportion

The test statistic is defined by:

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where:

$p = \frac{m}{n}$ distinguished proportion for the sample taken from the population,

m – frequency of values distinguished in the sample,

n – sample size.

The test statistic with a continuity correction is defined by:

$$Z = \frac{|p - p_0| - \frac{1}{2n}}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

The Z statistic with and without a continuity correction asymptotically (for large sizes) has the **normal distribution**.

Binominal test for one proportion

The binominal test for one proportion uses directly the **binominal distribution** which is also called the Bernoulli distribution, which belongs to the group of discrete distributions (such distributions, where the analysed variable takes in the finite number of values). The analysed variable can take in $k = 2$ values. The first one is usually defined with the name of a success and the other one with the name of a failure. The probability of occurrence of a success (distinguished probability) is p_0 , and a failure $1 - p_0$.

The probability for the specific point in this distribution is calculated using the formula:

$$P(m) = \binom{n}{m} p_0^m (1 - p_0)^{n-m},$$

where:

$$\binom{n}{m} = \frac{n!}{m!(n-m)!},$$

m – frequency of values distinguished in the sample,

n – sample size.

Based on the total of appropriate probabilities P a one-sided and a two-sided **p value** is calculated, and a two-sided p value is defined as a doubled value of the less of the one-sided probabilities.

The **p value** is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

Note

Note that, for the **estimator** from the sample, which in this case is the value of the p proportion, a **confidence interval** is calculated. The interval for a large sample size can be based on the **normal distribution** - so-called Wald intervals. The more universal are intervals proposed by Wilson (1927)[86] and by Agresti and Coull (1998)[2]. Clopper and Pearson (1934)[18] intervals are more adequate for small sample sizes.

Comparison of interval estimation methods of a binomial proportion was published by Brown L.D et al (2001)[15]

The settings window with the Z test for one proportion can be opened in Statistics menu → NonParametric tests (unordered categories) → Z for proportion.

Z for proportion

Statistical analysis : Z test for one proportion

▼ Frequency (numerator)

1-number of served dinners during one day
2-number of served dinners during one week
3-expected probability
4-day of the week

▼ Sample size (denominator)

1-number of served dinners during one day
2-number of served dinners during one week
3-expected probability
4-day of the week

▼ Expected proportion

1-number of served dinners during one day
2-number of served dinners during one week
3-expected probability
4-day of the week

Data Filter

variable	condition	value
4-day of the week	=	Friday

CI Method

Clopper-Pearson (Binomial Exact)

Variable contains

☒ Frequency (numerator)
☐ Proportion

☒ Continuity correction

0,05 significance level

Report options

☐ Add analysed data
☒ Add graph

OK Close

EXAMPLE 10.2 cont. (*dinners.pqs file*)

Assume, that you would like to check if on Friday $\frac{1}{5}$ of all the dinners during the whole week are served. For the chosen sample $m = 20$, $n = 150$.

number of served dinners during one day	number of served dinners during one week	expected probability	day of the week
33	150	0.2	Monday
29	150	0.2	Tuesday
32	150	0.2	Wednesday
36	150	0.2	Thursday
20	150	0.2	Friday

Select the options of the analysis and activate a **filter** selecting the appropriate day of the week – Friday. If you do not activate the filter, no error will be generated, only statistics for given weekdays will be calculated.

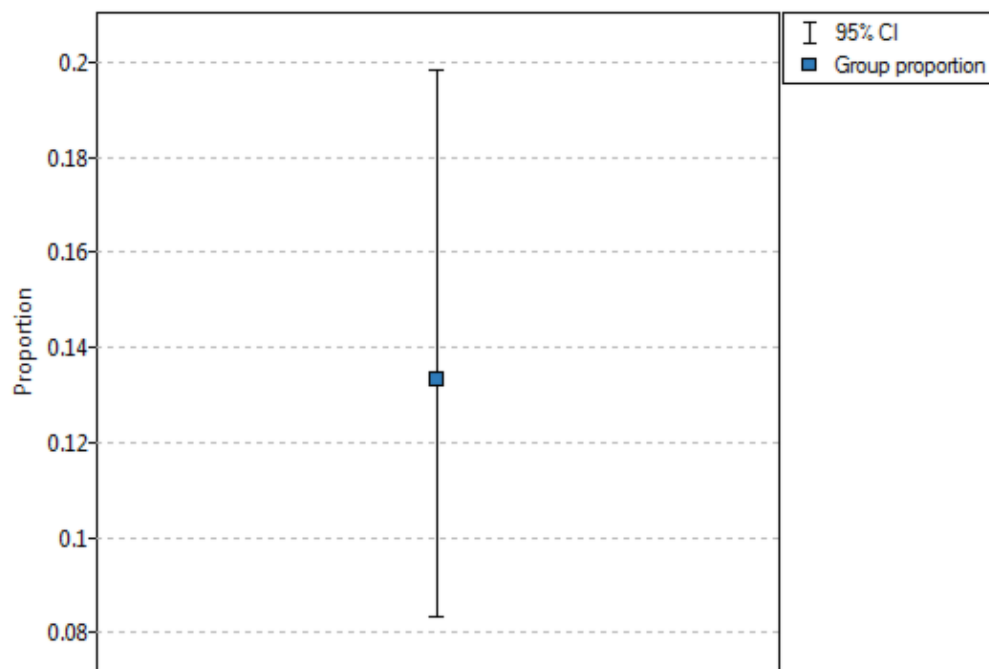
Hypotheses:

- \mathcal{H}_0 : on Friday, in a school canteen there are served $\frac{1}{5}$ out of all dinners which are served within a week,
- \mathcal{H}_1 : on Friday, in a school canteen there are significantly more than $\frac{1}{5}$ or less than $\frac{1}{5}$ dinners out of all the dinners served within a week in this canteen.

Z test for one proportion	
Analysis time	0,06sec.
Analysed variables	number of served dinners
Significance level	0,05
Continuity correction	No
Data Filter	day of the week=Friday
1	
Group proportion	0,133333
Clopper-Pearson (Binomial Exact)	
-95% CI for the proportion	0,083384
+95% CI for the proportion	0,198387
Z statistic	-2,041241
p-value (asymptotic)	0,041227
One sided p-value (exact)	0,022355
Two sided p-value (exact)	0,044711

Data:

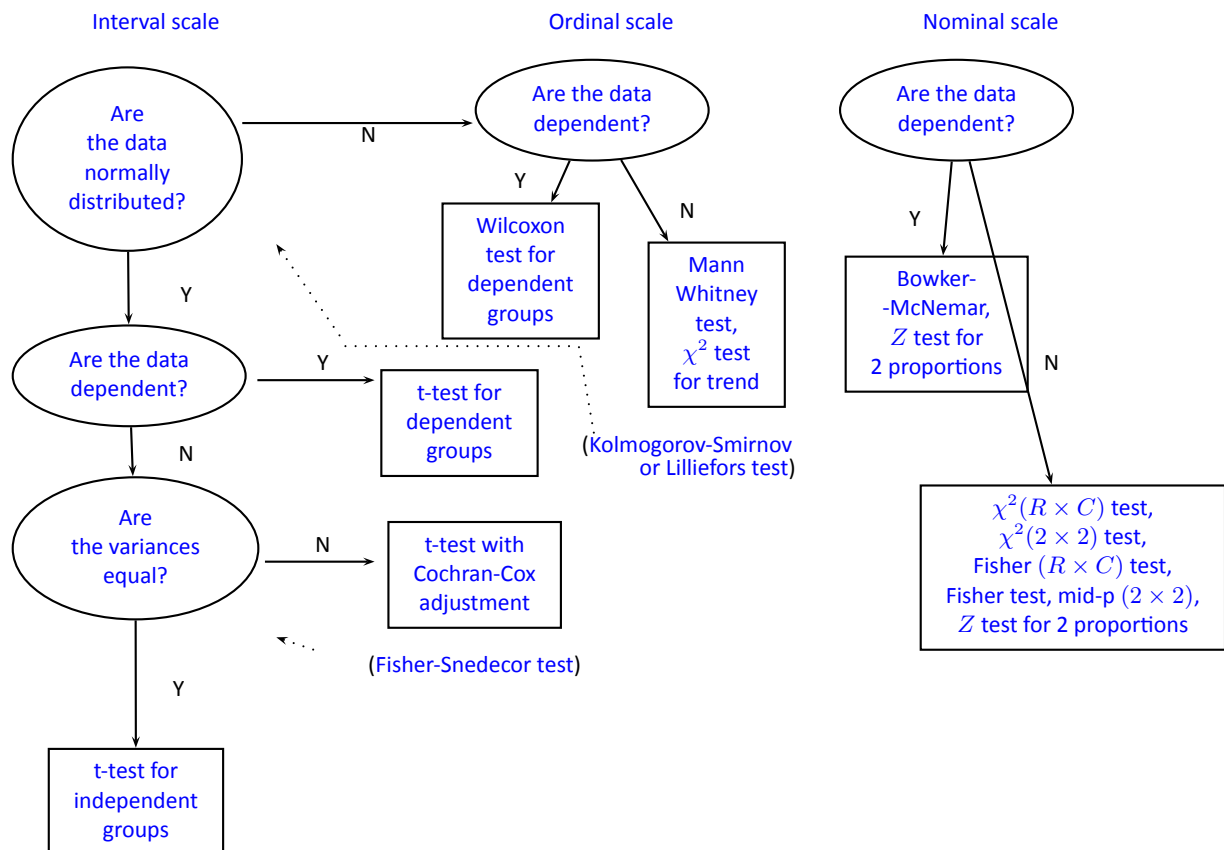
v.1	v.2	v.3
20	150	0,2



The proportion of the distinguished value in the sample is $p = \frac{m}{n} = 0.133$ and 95% Clopper-Pearson confidence interval for this fraction (0.083, 0.198) does not include the hypothetical value of 0.2.

Based on the Z test without the continuity correction (p value = 0.041227) and also on the basis of the exact value of the probability calculated from the binominal distribution (p value = 0.044711) you can assume (on the significance level $\alpha = 0.05$), that on Friday there are statistically less than $\frac{1}{5}$ dinners served within a week. However, after using the continuity correction it is not possible to reject the null hypothesis (p value = 0.052479).

11 COMPARISON - 2 GROUPS



11.1 PARAMETRIC TESTS

11.1.1 The Fisher-Snedecor test

The F-Snedecor test is based on a variable F which was formulated by Fisher (1924), and its distribution was described by Snedecor. This test is used to verify the hypothesis about equality of variances of an analysed variable for 2 populations.

Basic assumptions:

- measurement on an [interval scale](#),
- [normality of distribution](#) of an analysed feature in both populations,
- an [independent model](#).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : \sigma_1^2 &= \sigma_2^2, \\ \mathcal{H}_1 : \sigma_1^2 &\neq \sigma_2^2,\end{aligned}$$

where:

σ_1^2, σ_2^2 – variances of an analysed variable of the 1st and the 2nd population.

The test statistic is defined by:

$$F = \frac{sd_1^2}{sd_2^2},$$

where:

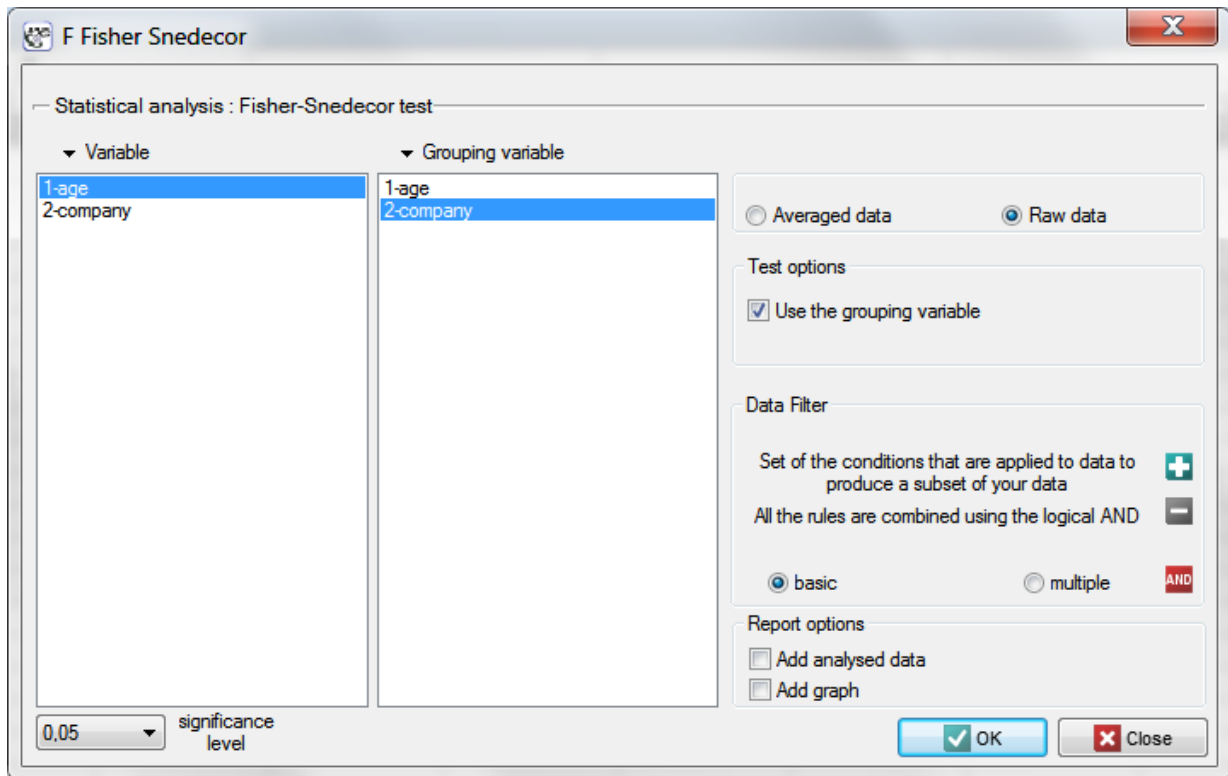
sd_1^2, sd_2^2 – variances of an analysed variable of the samples chosen randomly from the 1st and the 2nd population.

The test statistic has the [F Snedecor distribution](#) with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The settings window with the Fisher-Snedecor test can be opened in Statistics menu → Parametric tests → F Fisher Snedecor.



Note

Calculations can be based on [raw data](#) or data that are averaged like: arithmetic means, standard deviations and sample sizes.

11.1.2 The t-test for independent groups

The *t*-test for independent groups is used to verify the hypothesis about the equality of [means](#) of an analysed variable in 2 populations.

Basic assumptions:

- measurement on an [interval scale](#),
- [normality of distribution](#) of an analysed feature in both populations,
- an [independent model](#),
- [equality of variances](#) of an analysed variable in 2 populations.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \mu_1 = \mu_2, \\ \mathcal{H}_1 : & \mu_1 \neq \mu_2.\end{aligned}$$

where:

μ_1, μ_2 — means of an analysed variable of the 1st and the 2nd population.

The test statistic is defined by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 \cdot sd_1^2 + n_2 \cdot sd_2^2}{n_1 n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

where:

\bar{x}_1, \bar{x}_2 – means of an analysed variable of the 1st and the 2nd sample,

n_1, n_2 – the 1st and the 2nd sample size,

sd_1^2, sd_2^2 – **variances** of an analysed variable of the 1st and the 2nd sample.

The test statistic has the **t-Student distribution** with $df = n_1 + n_2 - 2$ degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 and accept \mathcal{H}_1 ,

if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

Note:

- **pooled standard deviation** is defined by:

$$SD_p = \sqrt{\frac{n_1 \cdot sd_1^2 + n_2 \cdot sd_2^2}{n_1 n_2 - 2}},$$

- **standard error of difference of means** is defined by:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{n_1 \cdot sd_1^2 + n_2 \cdot sd_2^2}{n_1 n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

11.1.3 The t-test with the Cochran-Cox adjustment

The Cochran-Cox adjustment relates to the **t-test for independent groups** (1957)[21] and is calculated when variances of analysed variables in both populations are different.

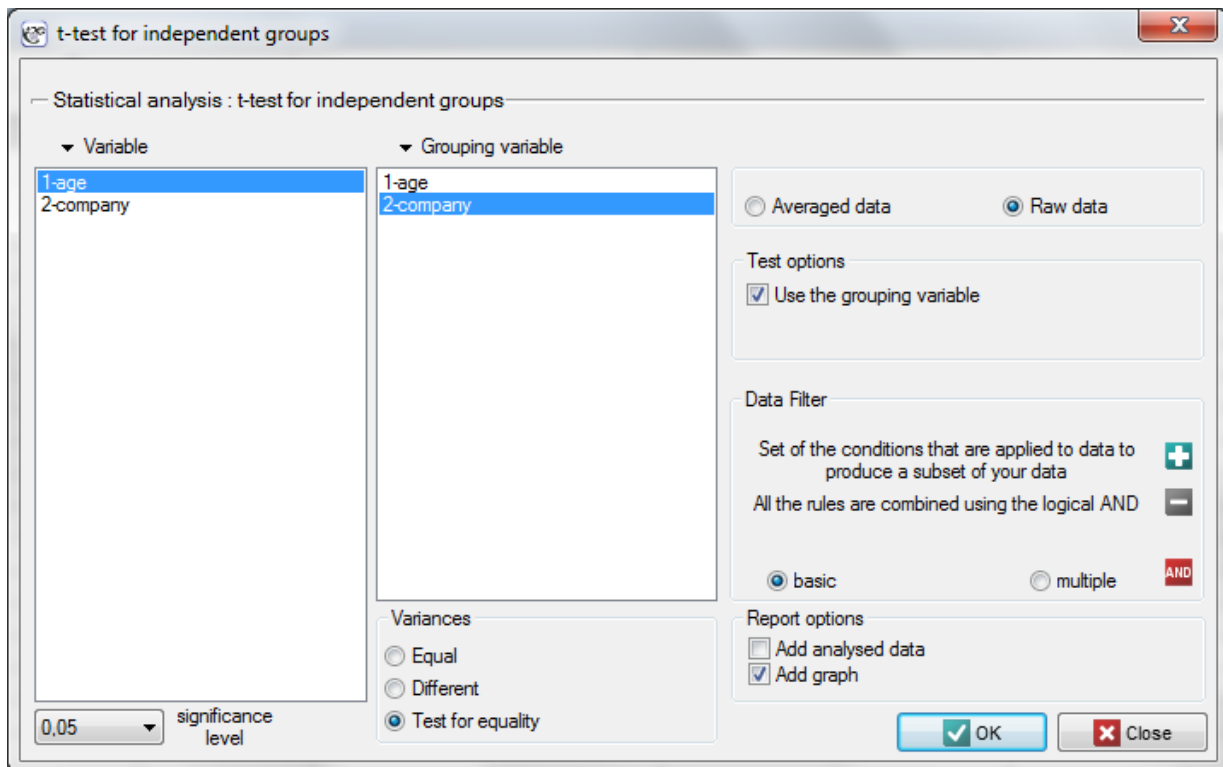
The test statistic is defined by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}.$$

The test statistic has the **t-Student distribution** with degrees of freedom proposed by Satterthwaite (1946)[73] and calculated using the formula:

$$df = \frac{\left(\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2} \right)^2}{\left(\frac{sd_1^2}{n_1} \right)^2 \cdot \frac{1}{(n_1-1)} + \left(\frac{sd_2^2}{n_2} \right)^2 \cdot \frac{1}{(n_2-1)}}.$$

The settings window with the **t- test for independent groups** can be opened in Statistics menu→Parametric tests→**t-test for independent groups** or in **Wizard**.



If, in the window which contains the options related to the variances, you have chosen:

- equal, the t -test for independent groups will be calculated ,
- different, the t -test with the Cochran-Cox adjustment will be calculated,
- check equality, to calculate the Fisher-Snedecor test, basing on its result and set the level of significance, the t -test for independent groups with or without the Cochran-Cox adjustment will be calculated.

Note

Calculations can be based on [raw data](#) or data that are averaged like: arithmetic means, standard deviations and sample sizes.

EXAMPLE 11.1. (age.pqs file)

There is an experiment, in which 100 people have been chosen randomly from the population of workers of 2 different transport companies. There are 50 people chosen from each company. Before the experiment begins, you should check if the average age of both companies workers is similar, because another step in the experiment depends on this. The age of each participant is written using years.

Age (company 1): 27, 33, 25, 32, 34, 38, 31, 34, 20, 30, 30, 27, 34, 32, 33, 25, 40, 35, 29, 20, 18, 28, 26, 22, 24, 24, 25, 28, 32, 32, 33, 32, 34, 27, 34, 27, 35, 28, 35, 34, 28, 29, 38, 26, 36, 31, 25, 35, 41, 37

Age (company 2): 38, 34, 33, 27, 36, 20, 37, 40, 27, 26, 40, 44, 36, 32, 26, 34, 27, 31, 36, 36, 25, 40, 27, 30, 36, 29, 32, 41, 49, 24, 36, 38, 18, 33, 30, 28, 27, 26, 42, 34, 24, 32, 36, 30, 37, 34, 33, 30, 44, 29

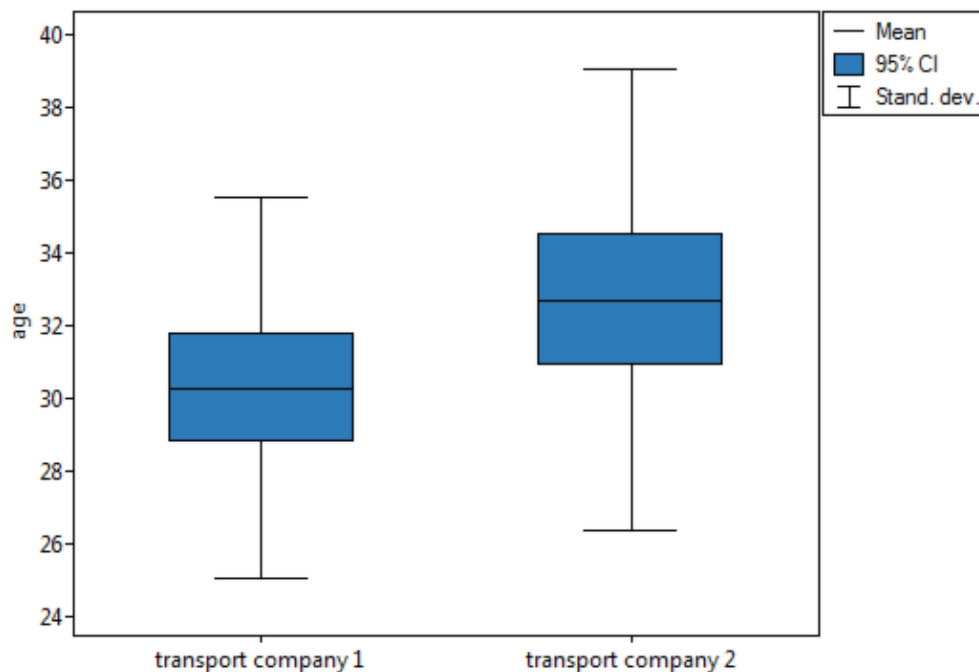
The age distribution in both groups is a normal one (it was tested with the [Lilliefors test](#)) with the mean of $\bar{x}_1 = 30.26$ and the standard deviation of $sd_1 = 5.23$ for the first group and $\bar{x}_2 = 32.68$ and $sd_2 = 6.36$ for the second group. The [Fisher-Snedecor test](#) also indicates that the variances of the

age in both companies are equal (p value = 0.176168). It means that all assumptions of the t -test for independent groups are fulfilled.

Hypotheses:

- \mathcal{H}_0 : the mean of the age of the first company workers is the same as the mean of the second company workers age,
 \mathcal{H}_1 : the mean of the age of the first company workers differs from the mean of the second company workers age.

t-test for independent groups	
Analysis time	0.03sec.
Analysed variables	age,company
Significance level	0.05
Correction for different variances	No
Grouping variable	company(transport comp
Group name	transport company 1
Group size	50
Group mean	30.26
Group standard deviation	5.23259
Group name	transport company 2
Group size	50
Group mean	32.68
Group standard deviation	6.358154
Difference of the means	-2.42
-95% CI for the difference	-4.730965
+95% CI for the difference	-0.109035
Std. err. of the difference	1.164527
Pooled standard deviation	5.822634
t-statistic	-2.078097
Degrees of freedom	98
two sided p-value	0.040314
Fisher-Snedecor test	
variance ratio F	0.677285
p-value	0.176168



If you compare the p value = 0.040314 with the significance level $\alpha = 0.05$ you draw the conclusion that the average age of all the workers chosen from both companies is different. The first company workers are a little bit more than 2 years younger than the second company workers.

11.1.4 The t-test for dependent groups

The t -test for dependent groups is used when the measurement of an analysed variable you do twice, each time in different conditions (but you should assume, that variances of the variable in both measurements are pretty close to each other). We want to check how big is the difference between the pairs of measurements ($d_i = x_{1i} - x_{2i}$). This difference is used to verify the hypothesis informing us that the **mean** of the difference in the analysed population is 0.

Basic assumptions:

- measurement on an **interval scale**,
- **normality of distribution** of measurements d_i (or the normal distribution for an analysed variable in each measurement),
- a **dependent model**.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : \mu_0 &= 0, \\ \mathcal{H}_1 : \mu_0 &\neq 0,\end{aligned}$$

where:

μ_0 , — mean of the differences d_i in a population.

The test statistic is defined by:

$$t = \frac{\bar{d}}{sd_d} \sqrt{n},$$

where:

\bar{d} – mean of differences d_i in a sample,

sd_d – **standard deviation** of differences d_i in a sample,

n – number of differences d_i in a sample.

Test statistic has the **t-Student distribution** with $n - 1$ degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 and accept \mathcal{H}_1 ,

if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

Note

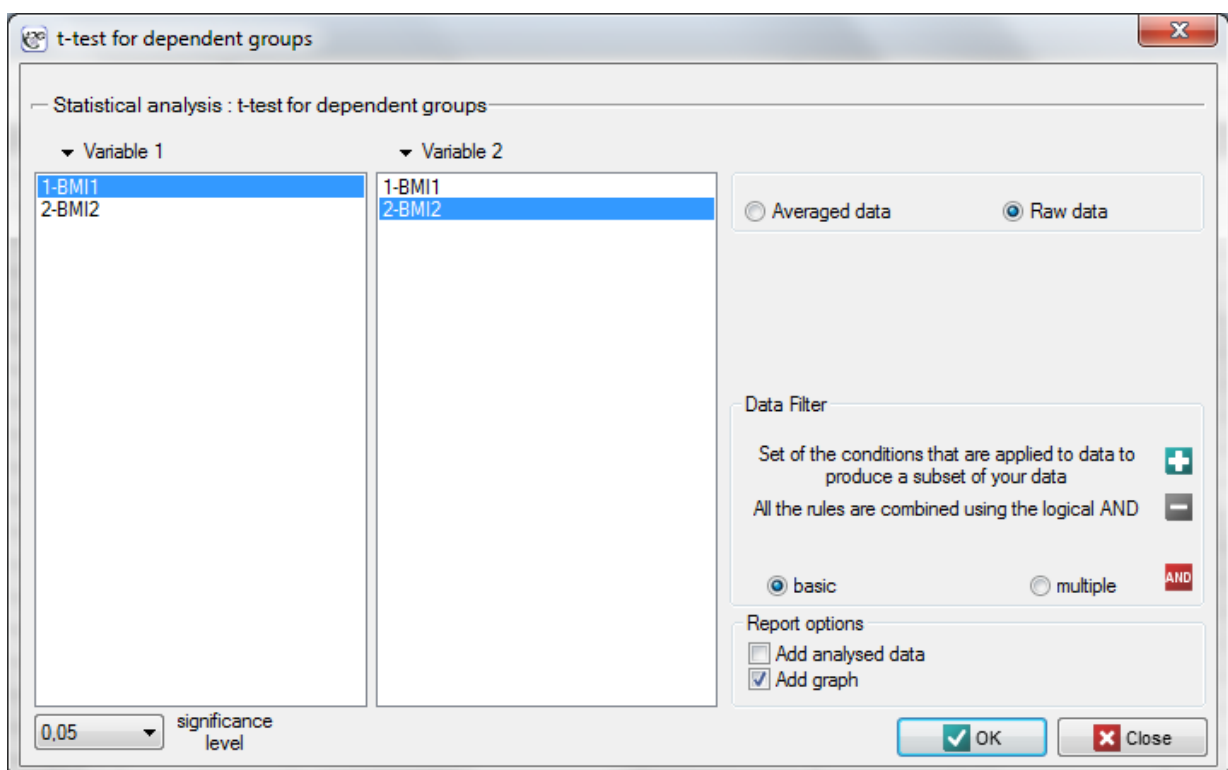
- **standard deviation of the difference** is defined by:

$$sd_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}},$$

- **standard error of the mean of differences** is defined by:

$$SEM_d = \frac{SD_d}{\sqrt{n}}.$$

The settings window with the *t*-test for dependent groups can be opened in Statistics menu → Parametric tests → *t*-test for dependent groups or in **Wizard**.



Note

Calculations can be based on **raw data** or data that are averaged like: arithmetic mean of difference, standard deviation of difference and sample size.

11.2 NONPARAMETRIC TESTS

11.2.1 The Mann-Whitney U test

The Mann-Whitney U test is also called as the Wilcoxon Mann-Whitney test (Mann and Whitney (1947)[55] and Wilcoxon (1949)[85]). This test is used to verify a hypothesis determining insignificance of differences between medians of an analysed variable in 2 populations (but you should assume that the distributions of a variable are pretty similar to each other).

Basic assumptions:

- measurement on an [ordinal scale](#) or on an [interval scale](#),
- an [independent model](#).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : \theta_1 &= \theta_2, \\ \mathcal{H}_1 : \theta_1 &\neq \theta_2,\end{aligned}$$

where:

θ_1, θ_2 medians of an analysed variable of the 1st and the 2nd population.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

Note

Depending on a sample size, the test statistic is calculated using by different formulas:

- For a small sample size:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1,$$

or

$$U' = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2,$$

where n_1, n_2 are sample sizes, R_1, R_2 are [rank](#) sums for the samples.

This statistic has the Mann-Whitney distribution and it does not contain any correction for ties. The value of the exact probability of the Mann-Whitney distribution is calculated with the accuracy up to the hundredth place of the fraction.

- For a large sample size:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum (t^3 - t)}{12(n_1 + n_2)(n_1 + n_2 - 1)}}},$$

where:

U can be replaced with U' ,

t – number of cases included in a [tie](#).

The formula for the Z statistic includes the correction for ties. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $\frac{n_1 n_2 \sum (t^3 - t)}{12(n_1 + n_2)(n_1 + n_2 - 1)} = 0$)

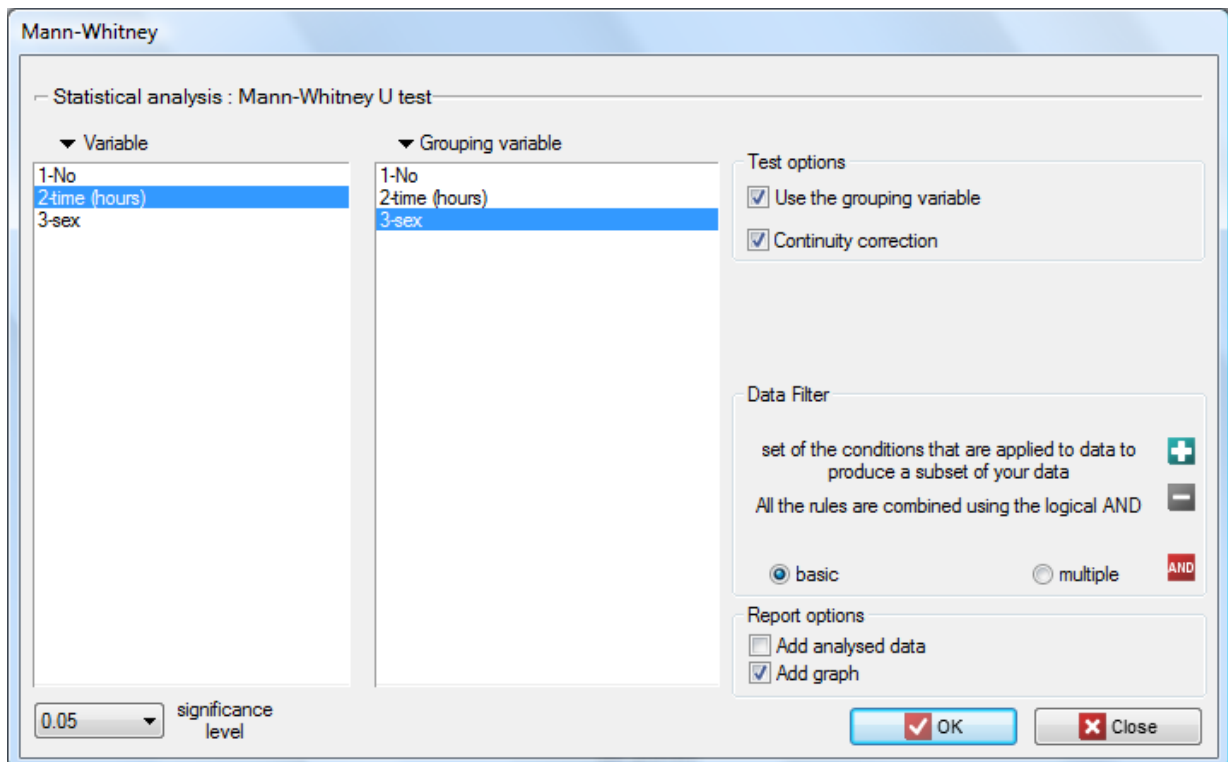
The Z statistic asymptotically (for large sample sizes) has the [normal distribution](#).

The Mann-Whitney test with the continuity correction (Marascuilo and McSweeney (1977)[60])

The continuity correction should be used to guarantee the possibility of taking in all the values of real numbers by the test statistic, according to the assumption of the normal distribution. The formula for the test statistic with the continuity correction is defined as:

$$Z = \frac{\left| U - \frac{n_1 n_2}{2} \right| - 0.5}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \sum (t^3 - t)}{12(n_1 + n_2)(n_1 + n_2 - 1)}}}.$$

The settings window with the Mann-Whitney U test can be opened in Statistics menu → NonParametric tests (ordered categories) → Mann-Whitney or in [Wizard](#).


EXAMPLE 11.2. (computer.pqs file)

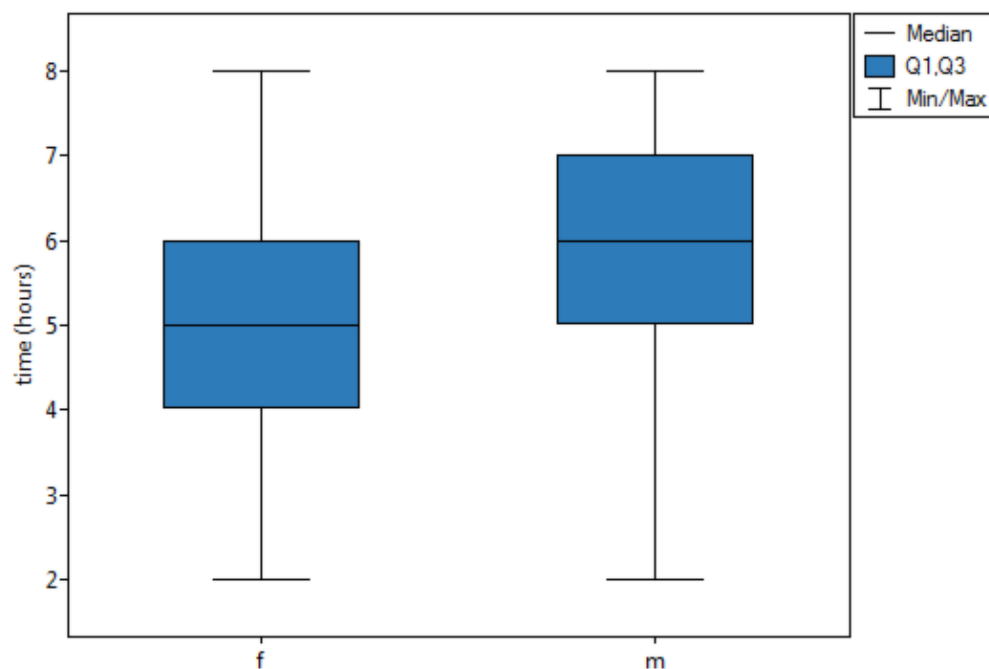
There was made a hypothesis that at some university male math students spend statistically more time in front of a computer screen than the female math students. To verify the hypothesis from the population of people who study math at this university, there was drawn a sample consisting of 54 people (25 women and 29 men). These persons were asked how many hours they spend in front of the computer screens daily. There were obtained the following results:

(time, sex): (2, k) (2, m) (2, m) (3, k) (3, k) (3, k) (3, k) (3, m) (3, m) (4, k) (4, k) (4, k) (4, k) (4, m) (4, m) (5, k) (5, k) (5, k) (5, k) (5, k) (5, k) (5, k) (5, k) (5, k) (5, m) (5, m) (5, m) (5, m) (6, k) (6, k) (6, k) (6, k) (6, k) (6, m) (6, m) (6, m) (6, m) (6, m) (6, m) (6, m) (6, m) (6, m) (6, m) (7, k) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (7, m) (8, k) (8, m) (8, m).

Hypotheses:

- \mathcal{H}_0 : the median of the time spent in front of a computer screen is exactly the same both in the male and the female population of students, at the analysed university,
- \mathcal{H}_1 : the median of the time spent in front of a computer screen is different among the male population and the female population of students, at the analysed university.

Mann-Whitney U test	
Analysis time	0.02sec.
Analysed variables	time (hours),sex
Significance level	0.05
Continuity correction	Yes
Grouping variable	sex(f;m)
Group name	f
Group size	25
Mean of the ranks for the group	22.02
Group sum of ranks	550.5
Group median	5
Group name	m
Group size	29
Mean of the ranks for the group	32.224138
Group sum of ranks	934.5
Group median	6
U statistic	225.5
U statistic	499.5
p-value (exact)	0.014948
Z statistic (adjusted for ties)	2.413028
p-value (asymptotic)	0.015821



Based on the assumed level $\alpha = 0.05$ and the Z statistic of the Mann-Whitney test without the continuity correction (p value = 0.015441) and with the continuity correction (p value = 0.015821), and also based on the exact U statistic (p value = 0.014948) you can assume that there are statistically significant differences among male and female students, if it goes about the time spent in front of a computer. These differences are, that female students spend less time in front of a computer than male students (the mean of the ranks for women is 22.02 (the median is 5) and it is much lower than the mean of the ranks for men, which is 32.22 (median is 6)).

11.2.2 The Wilcoxon test (matched-pairs)

The Wilcoxon matched-pairs test, is also called as the Wilcoxon test for dependent groups (Wilcoxon 1945[?], 1949[?]). It is used if the measurement of an analysed variable you do twice, each time in different conditions. It is the extension for the two dependent samples of the [Wilcoxon test \(signed-ranks\)](#) – designed for a one sample. We want to check how big is the difference between the pairs of measurements ($d_i = x_{1i} - x_{2i}$) for each of i analysed objects. This difference is used to verify the hypothesis determining that the [median](#) of the difference in the analysed population counts to 0.

Basic assumptions:

- measurement on an [ordinal scale](#) or on an [interval scale](#),
- a [dependent model](#).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : \theta_0 &= 0, \\ \mathcal{H}_1 : \theta_0 &\neq 0,\end{aligned}$$

where:

θ_0 – median of the differences d_i in a population.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

Note

Depending on the sample size, the test statistic is calculated by using different formulas:

- For small a sample size:

$$T = \min \left(\sum R_-, \sum R_+ \right),$$

where:

$\sum R_+$ – sums of positive [ranks](#),

$\sum R_-$ – sums of negative ranks.

This statistic has the Wilcoxon distribution and does not contain any correction for ties.

- For a large sample size

$$Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}},$$

where:

n – number of ranked signs (number of the ranks),

t – number of the cases included in a [tie](#).

The formula for the Z statistic includes the correction for ties. This correction is used, when the ties occur (if there are no ties, the correction is not calculated, because of $\frac{\sum t^3 - \sum t}{48} = 0$).

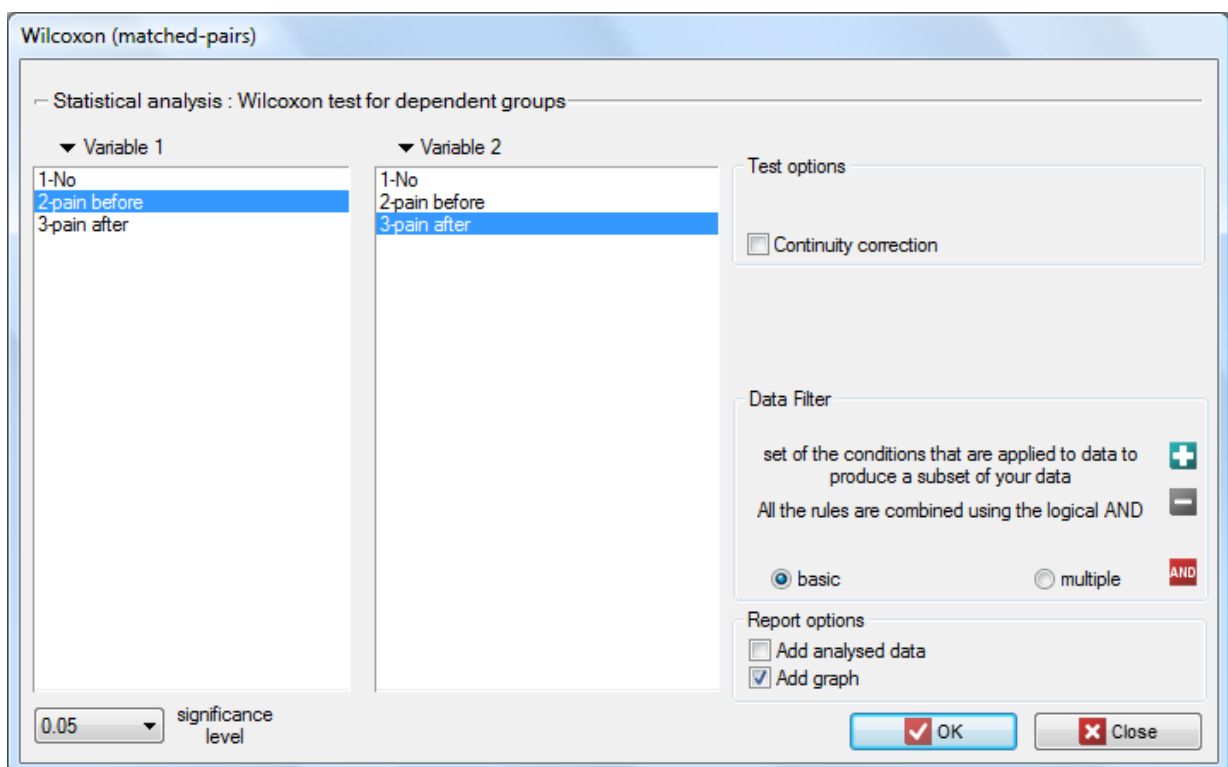
The Z statistic (for large sample sizes) asymptotically has the [normal distribution](#).

The Wilcoxon test with the continuity correction (Marascuilo and McSweeney (1977)[60])

The continuity correction is used to guarantee the possibility of taking in all the values of the real numbers by the test statistic, according to the assumption of the normal distribution. The test statistic with the continuity correction is defined by:

$$Z = \frac{\left| T - \frac{n(n+1)}{4} \right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}}.$$

The settings window with the Wilcoxon test for dependent groups can be opened in Statistics menu → NonParametric tests (ordered categories) → Wilcoxon (matched-pairs) or in [Wizard](#).


EXAMPLE 11.3. (pain.pqs file)

There was chosen a sample consisting of 22 patients suffering from a cancer. They were examined to check the level of felt pain (1 – 10 scale, where 1 means the lack of pain and 10 means unbearable pain). This examination was repeated after a month of the treatment with a new medicine which was supposed to lower the level of felt pain. There were obtained the following results:

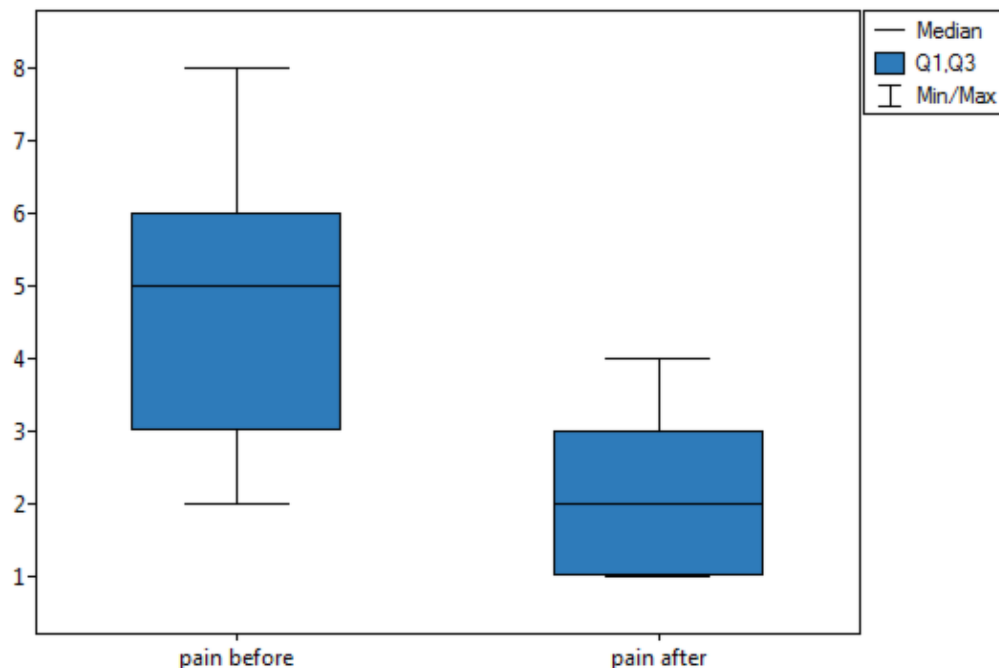
(pain before, pain after): (2, 2) (2, 3) (3, 1) (3, 1) (3, 2) (3, 2) (3, 3) (4, 1) (4, 3) (4, 4) (5, 1) (5, 1) (5, 2) (5, 4) (5, 4) (6, 1) (6, 3) (7, 2) (7, 4) (7, 4) (8, 1) (8, 3).

Now, you want to check if this treatment has any influence on the level of felt pain in the population, from which the sample was chosen.

Hypotheses:

- \mathcal{H}_0 : the median of the differences between the level of pain before and after a month of treatment in the analysed population comes to 0,
- \mathcal{H}_1 : the median of the differences between the level of pain before and after a month of treatment in the analysed population is different from 0.

Wilcoxon test for dependent groups	
Analysis time	0.02sec.
Analysed variables	pain before,pain after
Significance level	0.05
Continuity correction	Yes
Size = number of pairs	22
Count of omitted pairs (equal values)	3
Median of the difference	4
Sum of negative ranks	3.5
Sum of positive ranks	186.5
t statistic	3.5
p-value (exact)	0.0001
Z statistic (adjusted for ties)	3.68486
p-value (asymptotic)	0.000229



Comparing the p value = 0.0001 of the Wilcoxon test, based on the T statistic, with the significance level $\alpha = 0.05$ you assume, that there is a statistically significant difference if concerning the level of felt pain between these 2 examinations. The difference is, that the level of pain decreased (the sum of the negative ranks is significantly greater than the sum of the positive ranks). Exactly the same decision you would make on the basis of p value = 0.00021 or p value = 0.00023 of the Wilcoxon test which is based on the Z statistic or the Z statistic with the continuity correction.

11.2.3 TESTS FOR CONTINGENCY TABLES

Tests for contingency tables can be calculated on the basis of the data gathered as [contingency tables](#) or in the form of a [raw data](#). But there is also a possibility to [transform](#) the data from the contingency table to the raw form, or inversely.

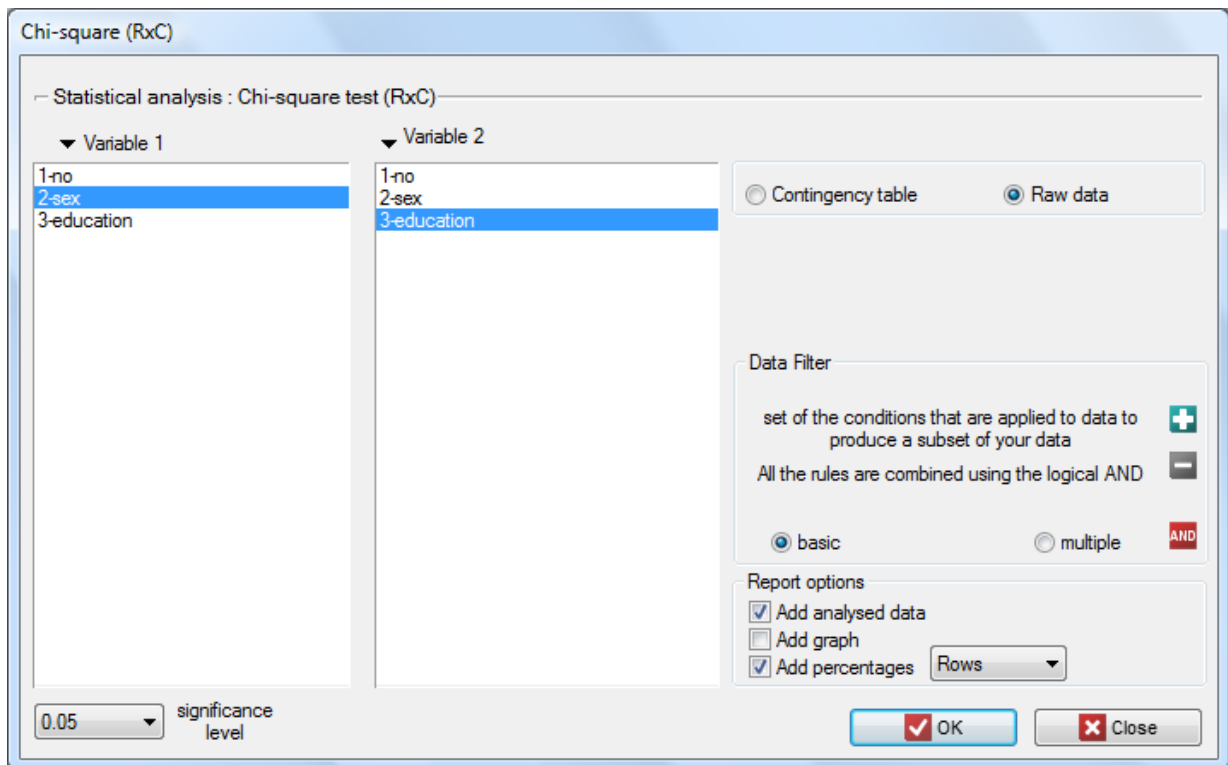
In the PQStat application there is a group of tests, which can be used on the base of one form as well as the other one. There are:

- The χ^2 test for the trend for $R \times 2$ tables,
- The χ^2 test and the Fisher test for $R \times C$ tables,
- The χ^2 test and the Fisher test for 2×2 tables and their corrections,
- The McNemar test, the Bowker test of the internal symmetry,
- The Test of significance for Cohen's Kappa.

EXAMPLE 11.4. (sex-education.pqs file)

There is a sample which consists of 34 persons ($n = 34$). You need to analyse the 2 features of these persons (X =sex, Y =education). Sex occurs in 2 categories (X_1 =woman, X_2 =man), education occurs in 3 categories, (Y_1 =primary+vocational Y_2 =secondary, Y_3 =higher).

In case of the raw data, when you open the window with the options for the test, for example the χ^2 test for $C \times R$ table, the option – raw data will be automatically selected.



In case of the data gathered in a contingency table, it is worth to select this data (the values numbers without headings) before you open the above-mentioned window. Doing it and opening the window, the contingency table will be automatically selected and all the data from the selection will be shown to you.

Chi-square (RxC)

Statistical analysis : Chi-square test (RxC)

fill with saved selection

	A	B	C
1	4	7	4
2	8	6	5

☒ Contingency table
 ☐ Raw data

Report options

☒ Add analysed data
☐ Add graph
☒ Add percentages

Rows

0.05 significance level

OK Close

In the test window, you can always change the default settings relating to the kind of the data organisation. In this window, you can also write the data which are supposed to be put into the contingency table.

As a result, you can return to the report, not only the test statistic and a p value, but also:

- **The contingency tables of observed frequencies** – data in the form of a contingency table. This table shows the distribution of observations for several features (several variables). The table of the 2 features (X , Y) – one of them has r possible categories and the other one c possible categories – is shown below (table(11.1)).

Table 11.1. The contingency table of $r \times c$ observed frequencies

Observed frequencies O_{ij}		Feature Y				
		Y_1	Y_2	...	Y_c	Total
Feature X	X_1	O_{11}	O_{12}	...	O_{1c}	$\sum_{j=1}^c O_{1j}$
	X_2	O_{21}	O_{22}	...	O_{2c}	$\sum_{j=1}^c O_{2j}$

	X_r	O_{r1}	O_{r2}	...	O_{rc}	$\sum_{j=1}^c O_{rj}$
	Total	$\sum_{i=1}^r O_{i1}$	$\sum_{i=1}^r O_{i2}$...	$\sum_{i=1}^r O_{ic}$	$n = \sum_{i=1}^r \sum_{j=1}^c O_{ij}$

Observed frequencies O_{ij} ($i = 1, 2, \dots, r$; $j = 1, 2, \dots, c$) show the frequencies of occurrence of all the particular categories for both features.

To return the table to the report, you should choose the option – add analysed data. For data from the example (11.4) the contingency table of the observed frequencies looks like this:

Data:	higher primary+ secondary		
female	7	4	4
male	6	8	5

- **The contingency table of expected frequencies** – for each contingency table of observed frequencies, can be created an adequate table of **expected frequencies**: E_{ij} (table(11.2)).

Table 11.2. The contingency table of $r \times c$ expected frequencies

Expected frequencies E_{ij}		Feature Y			
		Y_1	Y_2	...	Y_c
Feature X	X_1	E_{11}	E_{12}	...	E_{1c}
	X_2	E_{21}	E_{22}	...	E_{2c}

	X_r	E_{r1}	E_{r2}	...	E_{rc}

where:

$$E_{11} = \frac{\sum_{i=1}^r O_{i1} \times \sum_{j=1}^c O_{1j}}{n}, E_{12} = \frac{\sum_{i=1}^r O_{i2} \times \sum_{j=1}^c O_{1j}}{n}, E_{1c} = \frac{\sum_{i=1}^r O_{ic} \times \sum_{j=1}^c O_{1j}}{n}$$

$$E_{21} = \frac{\sum_{i=1}^r O_{i1} \times \sum_{j=1}^c O_{2j}}{n}, E_{22} = \frac{\sum_{i=1}^r O_{i2} \times \sum_{j=1}^c O_{2j}}{n}, E_{2c} = \frac{\sum_{i=1}^r O_{ic} \times \sum_{j=1}^c O_{2j}}{n}$$

$$E_{r1} = \frac{\sum_{i=1}^r O_{i1} \times \sum_{j=1}^c O_{rj}}{n}, E_{r2} = \frac{\sum_{i=1}^r O_{i2} \times \sum_{j=1}^c O_{rj}}{n}, E_{rc} = \frac{\sum_{i=1}^r O_{ic} \times \sum_{j=1}^c O_{rj}}{n}.$$

For the data from the example (11.4), the contingency table of expected frequencies looks like this:

Expected:	higher primary+ secondary		
female	5.74	5.29	3.97
male	7.26	6.71	5.03

- **The contingency table of percentages calculated from the sum of columns.** For the data from the example (11.4), the contingency table looks like this:

Columns:	higher primary+ secondary		
female	53.85%	33.33%	44.44%
male	46.15%	66.67%	55.56%

- **The contingency table of percentages calculated from the sum of rows.** For the data from the example (11.4), the contingency table looks like this:

Rows:	higher primary+ secondary		
female	46.67%	26.67%	26.67%
male	31.58%	42.11%	26.32%

- The contingency table of the percentages calculated from the sum of rows and columns (from total). For the data from the example (11.4), the table looks like this:

Sums:			
	higher	primary+secondary	
female	20.59%	11.76%	11.76%
male	17.65%	23.53%	14.71%

We can distinguish 2 approaches for analysed contingency tables. We can analyse the independence between both features or their homogeneities. It means to check if there are any differences between distribution of the first feature (variable) and the second one. However, these approaches sound differently, as they both lead to the same calculations.

11.2.4 The Chi-square test for trend for $R \times 2$ tables

The χ^2 test for trend is used to determine whether there is a trend in proportion for particular categories of an analysed variables (features). It is based on the data gathered in the [contingency tables of 2 features](#). The first feature has the possible r ordered categories: X_1, X_2, \dots, X_r and the second one has 2 categories: G_1, G_2 (table (11.3)).

Table 11.3. The contingency table of $r \times 2$ observed frequencies

Observed frequencies O_{ij}		Feature 2 (group)		
		G_1	G_2	Total
Feature 1 (feature X)	X_1	O_{11}	O_{12}	$W_1 = O_{11} + O_{12}$
	X_2	O_{21}	O_{22}	$W_2 = O_{21} + O_{22}$

	X_r	O_{r1}	O_{r2}	$W_r = O_{r1} + O_{r2}$
	Total	$C_1 = \sum_{i=1}^r O_{i1}$	$C_2 = \sum_{i=1}^r O_{i2}$	$n = C_1 + C_2$

Basic assumptions:

- measurement on an [ordinal scale](#) or on an [interval scale](#),
- an [independent model](#) (the second feature – 2 independent groups).

Hypotheses:

\mathcal{H}_0 : In the analysed population the trend in a proportion of p_1, p_2, \dots, p_r does not exist,

\mathcal{H}_1 : There is the trend in a proportion of p_1, p_2, \dots, p_r in the analysed population.

where:

p_1, p_2, \dots, p_r are the proportions $p_1 = \frac{O_{11}}{W_1}, p_2 = \frac{O_{21}}{W_2}, \dots, p_r = \frac{O_{r1}}{W_r}$.

The test statistic is defined by:

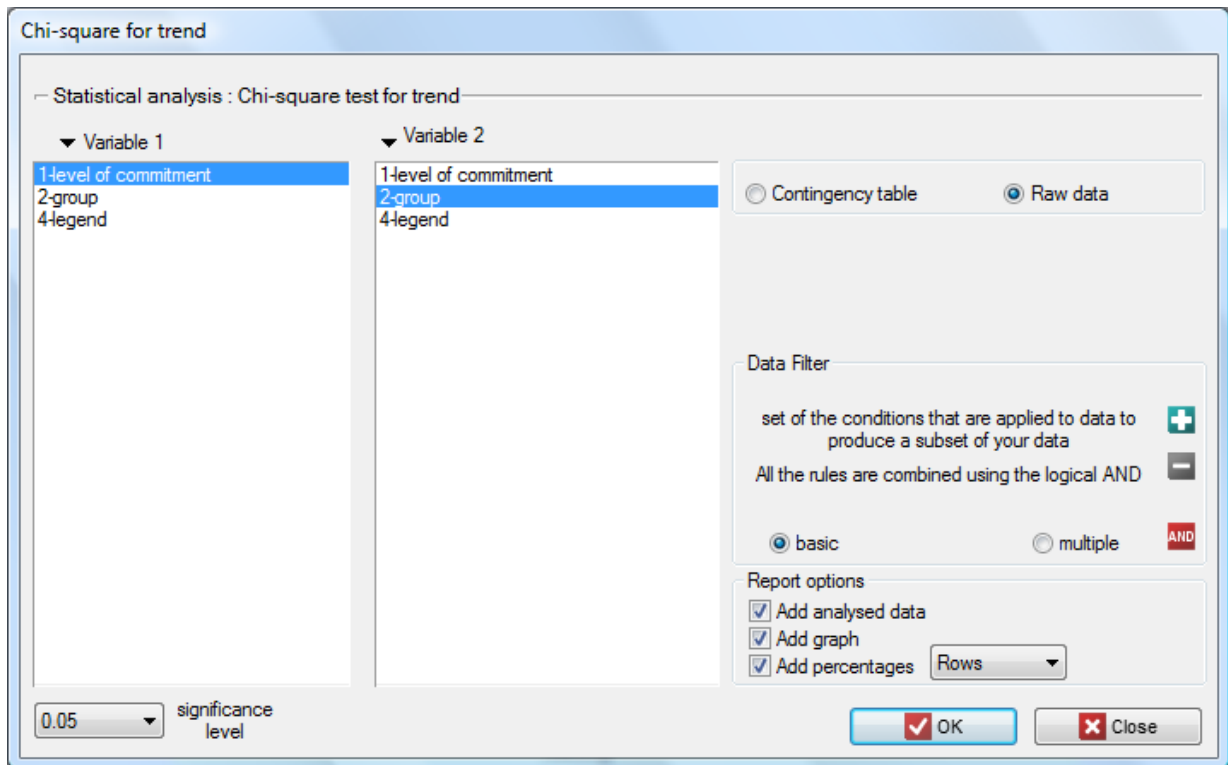
$$\chi^2 = \frac{\left[\left(\sum_{i=1}^r i \cdot O_{i1} \right) - C_1 \left(\sum_{i=1}^r \frac{i \cdot W_i}{n} \right) \right]^2}{\frac{C_1}{n} \left(1 - \frac{C_1}{n} \right) \left[\left(\sum_{i=1}^n i^2 W_i \right) - n \left(\sum_{i=1}^n \frac{i \cdot W_i}{n} \right)^2 \right]}.$$

This statistic asymptotically (for large expected frequencies) has the [\$\chi^2\$ distribution](#) with 1 degree of freedom.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

The settings window with the Chi-square test for trend can be opened in Statistics menu \rightarrow NonParametric tests (ordered categories) \rightarrow Chi-square for trend or in [Wizard](#).



EXAMPLE 11.5. (viewers.pqs file)

Because of the decrease in people watching some particular soap opera there was carried out an opinion survey. 100 persons were asked, who has recently started watching this soap opera, and 300 persons were asked, who has watched it regularly from the beginning. They were asked about the level of preoccupation with the character's life. The results are written down in the table below:

Level of commitment	grupa		
	group of new viewers	group of steady viewers	total
rather small	7	7	14
average	13	25	38
rather high	30	58	88
high	24	99	123
very high	26	111	137
total	100	300	400

The new viewers consist of 25% of all the analysed viewers. This proportion is not the same for each level of commitment, but looks like this:

Level of commitment	group		
	group of new viewers	group of steady viewers	total
rather small	$p_1=50.00\%$	50.00%	100%
average	$p_2=34.21\%$	65.79%	100%
rather high	$p_3=34.09\%$	65.91%	100%
high	$p_4=19.51\%$	80.49%	100%
very high	$p_5=18.98\%$	81.02%	100%
total	25.00%	75.00%	100%

Hypotheses:

\mathcal{H}_0 : in the population of the soap opera viewers, the trend in proportions of p_1, p_2, p_3, p_4, p_5 does not exist,

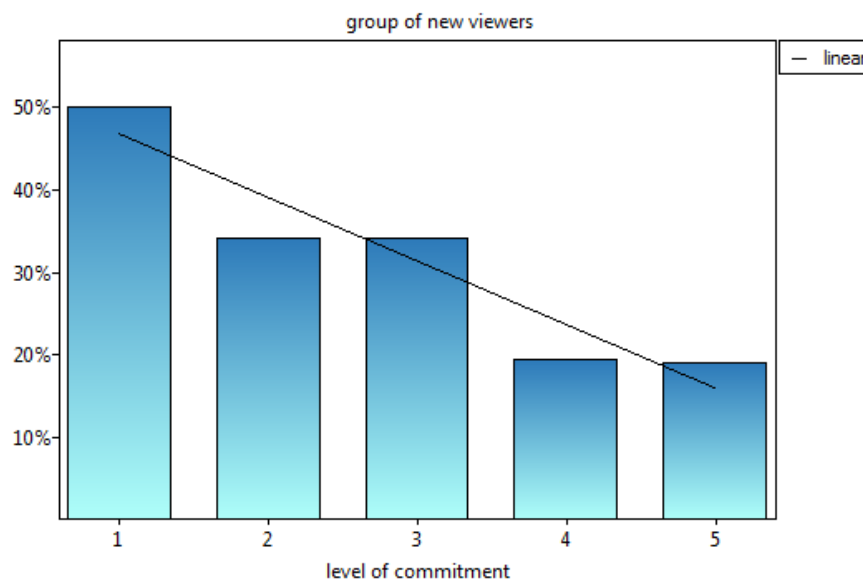
\mathcal{H}_1 : in the population of the soap opera viewers, the trend in proportions of p_1, p_2, p_3, p_4, p_5 does exist.

Chi-square test for trend	
Analysis time	0.04sec.
Analysed variables	level of commitment;group
Significance level	0.05
Group size1	100
Group size2	300
Chi-square statistic	12.3702523
Degrees of freedom	1
p-value	0.0004362

Expected:		
	group of n	group of s
1	3.5	10.5
2	9.5	28.5
3	22	66
4	30.75	92.25
5	34.25	102.75

Data:		
	group of n	group of s
1	7	7
2	13	25
3	30	58
4	24	99
5	26	111

Rows:		
	group of n	group of s
1	50%	50%
2	34.21%	65.79%
3	34.09%	65.91%
4	19.51%	80.49%
5	18.98%	81.02%



The p value = 0.000436, compared with the significance $\alpha=0.05$ supports the alternative hypothesis informing that the trend in proportions of p_1, p_2, \dots, p_5 does exist. As shown in the contingency table of percentages calculated from the sum of columns, there is a decreasing trend (the more interested in the character's life the group of viewers is, the smaller part of the group of new viewers is).

11.2.5 The Chi-square test and Fisher test for $R \times C$ tables

These tests are based on the data gathered in the form of a contingency table of 2 features (X, Y). One of them has possible r categories X_1, X_2, \dots, X_r and the other one c categories Y_1, Y_2, \dots, Y_c (look at

the table (11.1)).

Basic assumptions:

- measurement on a **nominal scale** (alternatively: an **ordinal** or an **interval**),
- an **independent model**.

The additional assumption for the χ^2 :

- large **expected frequencies** (according to Cochran interpretation (1952)[20] none of these expected frequencies can be < 1 and no more than 20% of expected frequencies can be < 5).

- General hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & O_{ij} = E_{ij} \text{ for all categories,} \\ \mathcal{H}_1 : & O_{ij} \neq E_{ij} \text{ for at least one category,}\end{aligned}$$

where:

O_{ij} – **observed frequencies** in a contingency table,

E_{ij} – **expected frequencies** in a contingency table.

- Hypotheses in the meaning of independence:

\mathcal{H}_0 : there is no dependence between the analysed features of the population (both classifications are statistically independent according to X and Y feature),

\mathcal{H}_1 : there is a dependence between the analysed features of the population.

- Hypotheses in the meaning of homogeneity:

\mathcal{H}_0 : in the analysed population, the distribution of X feature categories is exactly the same for each category of Y feature,

\mathcal{H}_1 : in the analysed population distribution, the of X feature categories is different for at least one category of Y feature.

Compare the **p value**, calculated on the basis of the **test statistic**, with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha & \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The Chi-square test for $R \times C$ tables

The χ^2 test for $r \times c$ tables is also known as the Pearson's Chi-square test (Karl Pearson 1900). This test is an extension on 2 features of the **χ^2 test (goodness-of-fit)**.

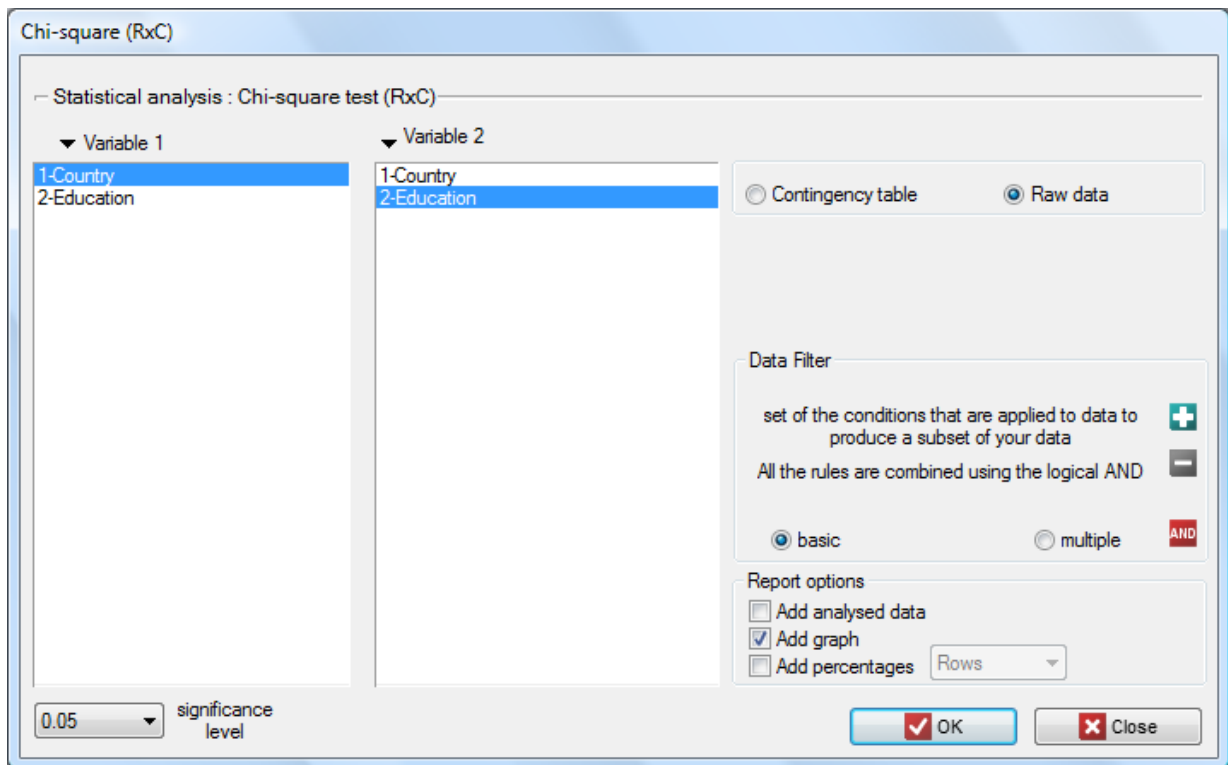
The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

This statistic asymptotically (for large expected frequencies) has the **χ^2 distribution** with a number of degrees of freedom calculated using the formula: $df = (r - 1)(c - 1)$.

Compare the **p value**, calculated on the basis of the **test statistic**, with the significance level α .

The settings window with the Chi-square test (RxC) can be opened in Statistics menu \rightarrow NonParametric tests (unordered categories) \rightarrow Chi-square (RxC) or in **Wizard**.

**EXAMPLE 11.6.** (country-education.pqs file)

There is a sample of 605 persons ($n = 605$), who had 2 features analysed for (X =country of residence, Y =education). The first feature occurs in 4 categories, and the second one in 3 categories (X_1 =Country 1, X_2 =Country 2, X_3 =Country 3, X_4 =Country 4, Y_1 =primary, Y_2 =secondary, Y_3 =higher). The data distribution is shown below, in the contingency table:

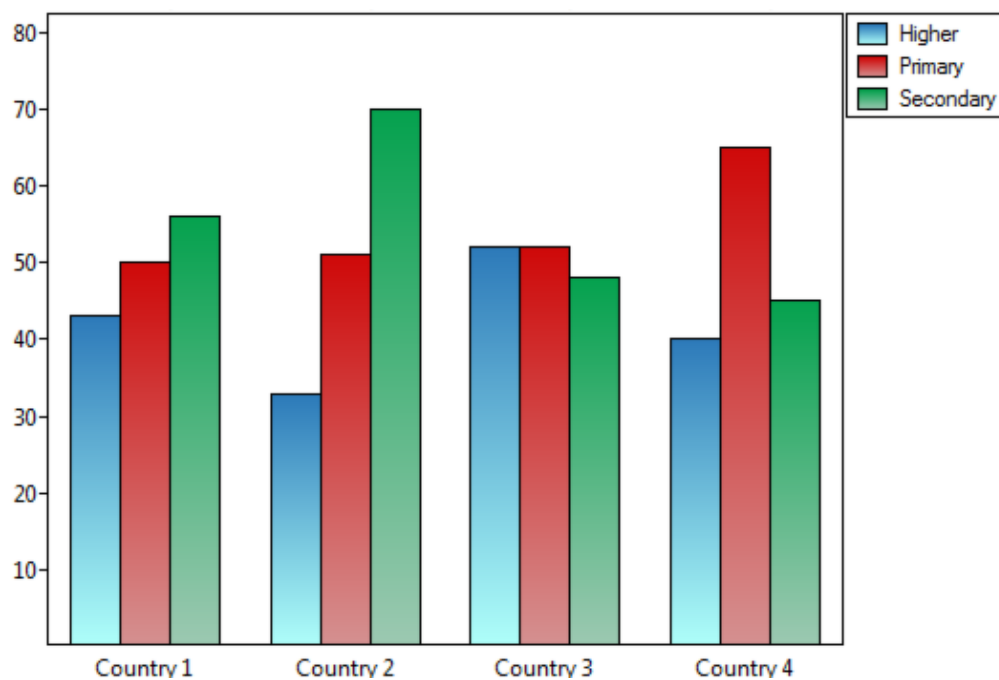
	Primary	Secondary	Higher
Country 1	50	56	43
Country 2	51	70	33
Country 3	52	48	52
Country 4	65	45	40

Based on this sample, you would like to find out if there is any dependence between education and country of residence in the analysed population.

Hypotheses:

- \mathcal{H}_0 : there is no dependence between education and country of residence in the analysed population,
 \mathcal{H}_1 : there is a dependence between education and country of residence in the analysed population.

Chi-square test (RxC)	
Analysis time	0.02sec.
Analysed variables	Country; Education
Significance level	0.05
Size	605
Chi-square statistic	13.817861
Degrees of freedom	6
p-value	0.031738



The table of the expected frequencies does not contain any values which are less than 5.

The p value = 0.03174. So, on the basis of the significance level $\alpha = 0.05$ we can draw the conclusion that there is a dependence between education and country of residence in the analysed population.

The Fisher test for $R \times C$ tables

The Fisher test for $r \times c$ tables is also called the Fisher-Freeman-Halton test (Freeman G.H., Halton J.H. (1951)[31]). This test is an extension on $r \times c$ tables of the [Fisher's exact test](#). It defines the exact probability of an occurrence specific distribution of numbers in the table (when we know n and we set the marginal totals).

If you define marginal sums of each row as:

$$W_i = \sum_{j=1}^c O_{ij},$$

where:

O_{ij} – [observed frequencies](#) in a table,

and the marginal sums of each column as:

$$K_i = \sum_{j=1}^r O_{ij}.$$

then, having defined the marginal sums for the different distributions of the observed frequencies represented by U_{ij} , you can calculate the P probabilities:

$$P = \frac{D^{-1} \prod_{j=1}^c K_j!}{U_{1j}! U_{2j}! \dots U_{rj}!},$$

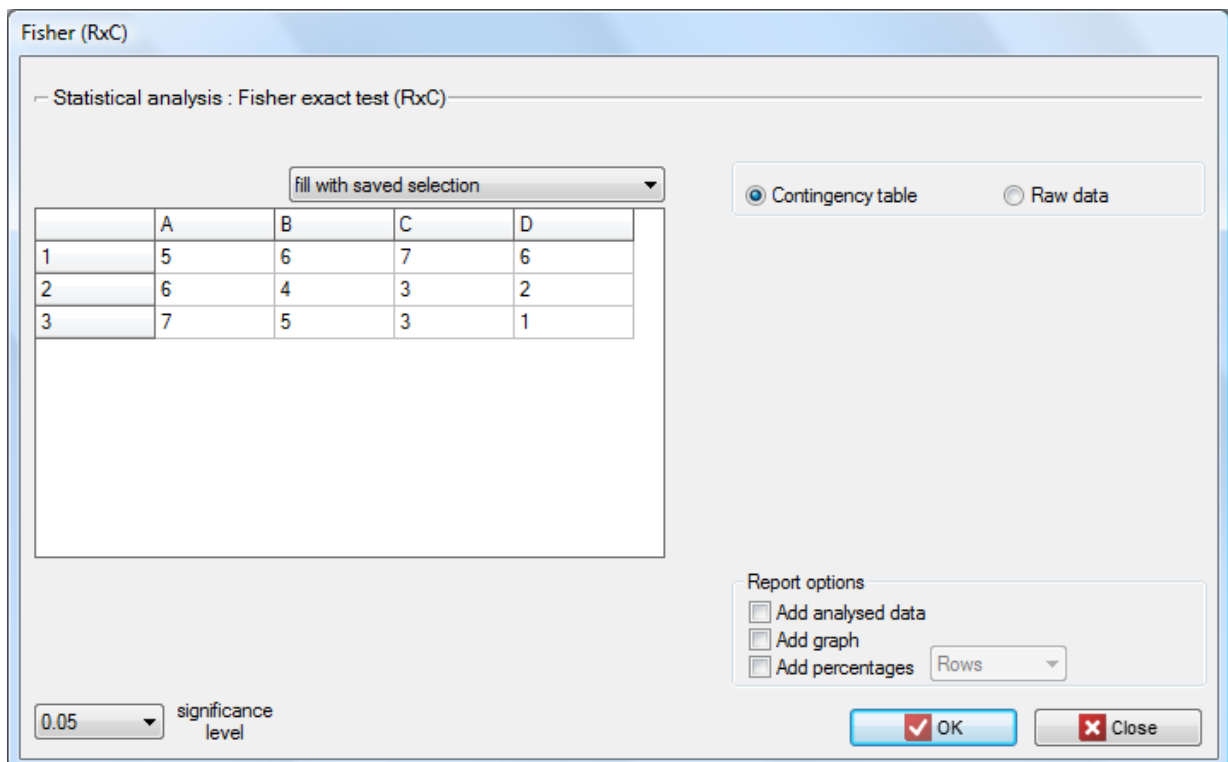
where

$$D = \frac{(W_1 + W_2 + \dots + W_r)!}{W_1! W_2! \dots W_r!}.$$

The exact significance level p : is the sum of P probabilities (calculated for new values U_{ij}), which are smaller or equal to P probability of the table with the initial numbers O_{ij} .

The exact p value is compared with the significance level α .

The settings window with the Fisher exact test (RxC) can be opened in Statistics menu → NonParametric tests (unordered categories) → Fisher (RxC) or in [Wizard](#).



Info.

The process of calculation of p values for this test is based on the algorithm published by Mehta (1986)[62].

Note

Note, that comparisons relating to 2 chosen categories can be made using the tests for contingency tables 2×2 and the Bonferroni correction [1].

11.2.6 The Chi-square test and the Fisher test for 2x2 tables (with corrections)

These tests are based on the data gathered in the form of a contingency table of 2 features (X, Y), each of them has 2 possible categories X_1, X_2 and Y_1, Y_2 (look at the table (11.1)).

Basic assumptions:

- measurement on a **nominal scale** (dichotomous variables — it means the variables of two categories),
- an **independent model**.

The additional assumption for the χ^2 test:

- large **expected frequencies** (according to the Cochran interpretation (1952)[20], none of these expected frequencies can be < 1 and no more than 20% of the expected frequencies can be < 5).

- General hypotheses:

$$\mathcal{H}_0 : O_{ij} = E_{ij} \text{ for all categories,}$$

$$\mathcal{H}_1 : O_{ij} \neq E_{ij} \text{ for at least one category,}$$

where:

O_{ij} — **observed frequencies** in a contingency table,

E_{ij} — **expected frequencies** in a contingency table.

- Hypotheses in the meaning of independence:

\mathcal{H}_0 : there is no dependence between the analysed features of the population (both classifications are statistically independent according to X and Y feature),

\mathcal{H}_1 : there is a dependence between the analysed features of the population.

- Hypotheses in the meaning of homogeneity:

\mathcal{H}_0 : in the analysed population, the distribution of X feature categories is exactly the same for both categories of Y feature,

\mathcal{H}_1 : in the analysed population, the distribution of X feature categories is different for both categories of Y feature.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\text{if } p \leq \alpha \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,$$

$$\text{if } p > \alpha \implies \text{there is no reason to reject } \mathcal{H}_0.$$

Note

Additionally for 2×2 contingency tables PQStat calculates also the **odds ratio** — **OR** and the **relative risk** — **RR** altogether with the **confidence intervals**. These intervals are calculated on the basis of the approximate χ^2 distribution — if they accompany the χ^2 test, or of the exact algorithms — if they accompany the Fisher's test and mid-p.

The Chi-square test for 2×2 tables

The χ^2 test for 2×2 tables — The Pearson's Chi-square test (Karl Pearson 1900) is constraint of the **χ^2 test for $r \times c$ tables**.

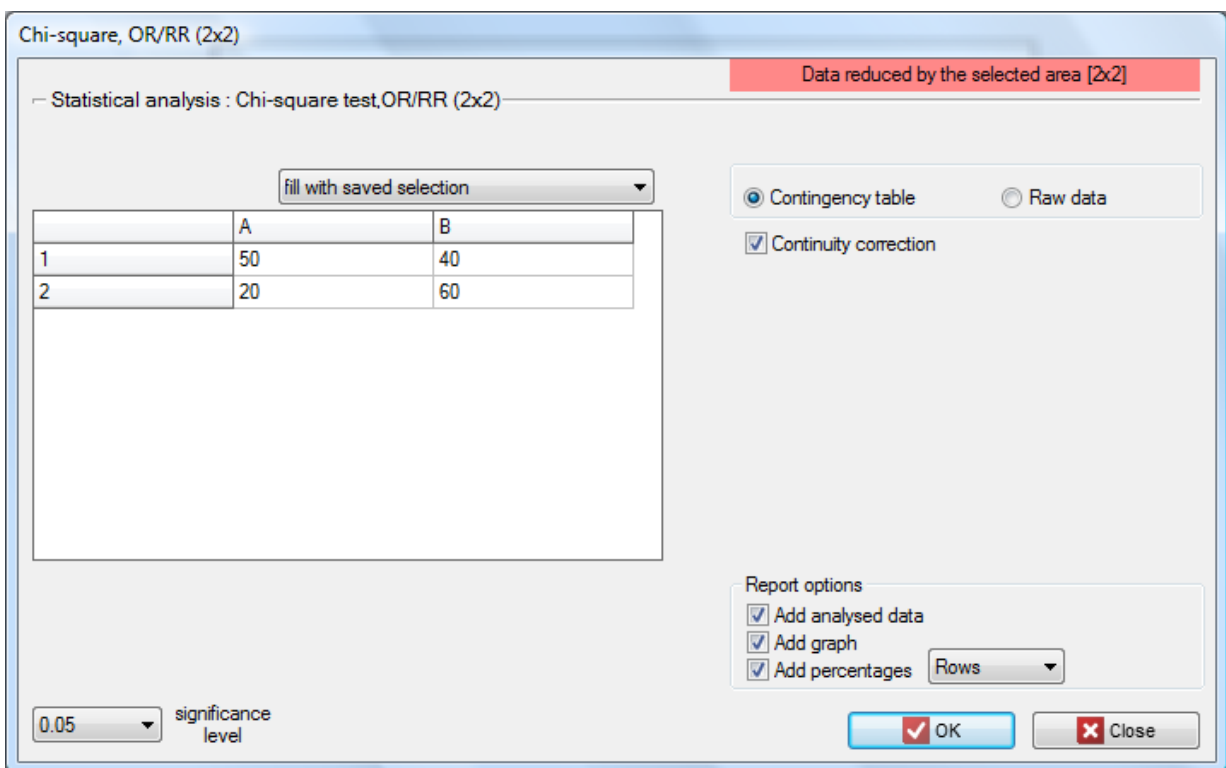
The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

This statistic asymptotically (for large expected frequencies) has the χ^2 distribution with a 1 degree of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α .

The settings window with the Chi-square test, OR/RR (2x2) can be opened in Statistics menu → NonParametric tests (unordered categories) → Chi-square, OR/RR (2x2) or in Wizard.



EXAMPLE 11.7. (sex-exam.pqs file)

There is a sample consisting of 170 persons ($n = 170$). Using this sample, you want to analyse 2 features (X =sex, Y =exam passing). Each of these features occurs in two categories ($X_1=f$, $X_2=m$, $Y_1=yes$, $Y_2=no$). Based on the sample you want to get to know, if there is any dependence between sex and exam passing in the above population. The data distribution is presented in the contingency table below:

Observed frequencies O_{ij}		exam passing		
		yes	no	total
sex	f	50	40	90
	m	20	60	80
	total	70	100	170

Hypotheses:

\mathcal{H}_0 : there is no dependence between sex and exam passing in the analysed population,

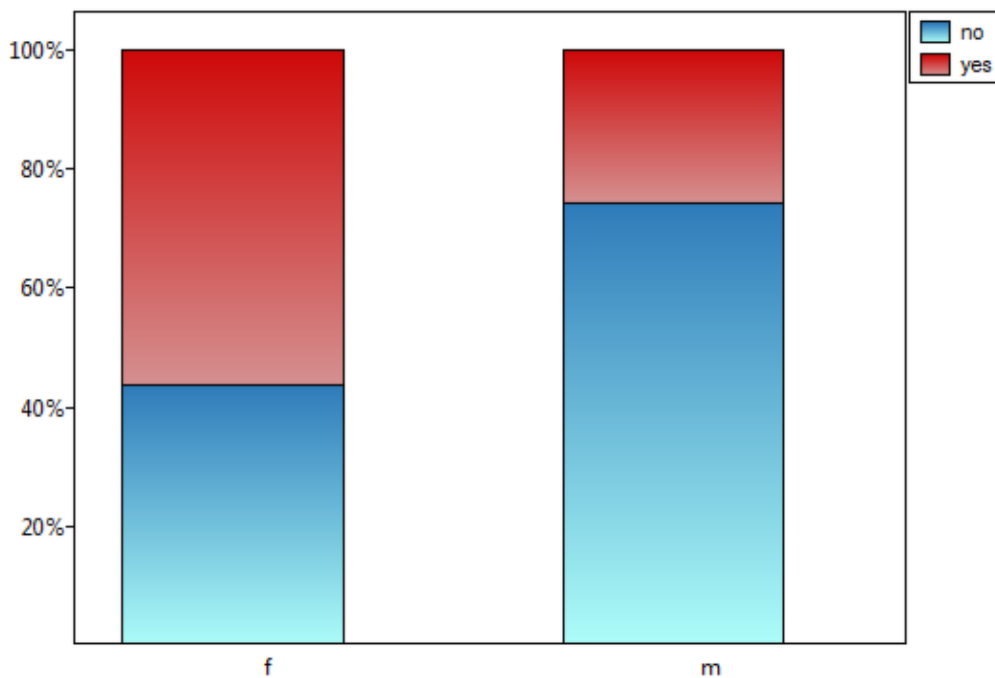
\mathcal{H}_1 : there is a dependence between sex and exam passing in the analysed population.

Chi-square test,OR/RR (2x2)	
Analysis time	0.02sec.
Analysed variables	Contingency table
Significance level	0.05
Size	170
Odds Ratio	3.75
-95% CI for the Odds Ratio	1.948
+95% CI for the Odds Ratio	7.218942
Statistic for the Odds Ratio	3.955391
p-value	0.000076
Relative Risk	2.222222
-95% CI for the Relative Risk	1.456985
+95% CI for the Relative Risk	3.389378
Statistic for the Relative Risk	3.707423
p-value	0.000209
Chi-square statistic	16.325397
Degrees of freedom	1
p-value	0.000053
Statistic with Yates correction	15.088259
Degrees of freedom	1
p-value with Yates correct.	0.000103

Expected:	
37.06	52.94
32.94	47.06

Data:	
50	40
20	60

Rows:	
55.56%	44.44%
25%	75%



The expected frequency table does not contain any values less than 5.

The p value = 0.000053. So, on the significance level $\alpha = 0.05$ we can accept the alternative hypothesis informing us that there is a dependence between sex and exam passing in the analysed population. Significantly, the exam is passed more often by women ($\frac{50}{90} = 55.56\%$ out of all the women in the sample

who passed the exam) than by men ($\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam).

The Chi-square test with the Yate's correction for continuity

The χ^2 test with the Yate's correction (Frank Yates (1934)[87]) is a more conservative test than the χ^2 test (it rejects a null hypothesis more rarely than the χ^2 test). The correction for continuity guarantees the possibility of taking in all the values of real numbers by a test statistic, according to the χ^2 distribution assumption.

The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}.$$

EXAMPLE (11.7) cont. (*sex-exam.pqs* file)

The p value for the χ^2 test with the Yate's correction is 0.000103. Similarly to the χ^2 test without the correction, on the significance level $\alpha = 0.05$, the alternative hypothesis can be accepted. The alternative hypothesis informs, that there is a dependence between sex and exam passing in the analysed population. Significantly, the exam was passed more often by women ($\frac{50}{90} = 55.56\%$ out of all the women in the sample who passed the exam) than by men ($\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam).

The Fisher test for 2×2 tables

The Fisher test for 2×2 tables is also called the Fisher exact test (R. A. Fisher (1934)[27], (1935)[28]). This test enables you to calculate the exact probability of the occurrence of the particular number distribution in a table (knowing n and defined marginal sums).

$$P = \frac{\binom{O_{11}+O_{21}}{O_{11}} \binom{O_{12}+O_{22}}{O_{12}}}{\binom{O_{11}+O_{12}+O_{21}+O_{22}}{O_{11}+O_{12}}}.$$

If you know each marginal sum, you can calculate the P probability for various configurations of observed frequencies. The exact p significance level is the sum of probabilities which are less or equal to the analysed probability.

The p value is compared with the significance level α .

The settings window with the Fisher exact test, mid-p (2x2) can be opened in Statistics menu → NonParametric tests (unordered categories) → Fisher, mid-p (2x2) or in Wizard.

Fisher, Mid-P (2x2)

Statistical analysis : Fisher exact test, mid-p (2x2)

Variable 1

- 1-No
- 2-sex
- 3-exam passed

Variable 2

- 1-No
- 2-sex
- 3-exam passed

☐ Contingency table ☒ Raw data

Data Filter

set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

☒ basic ☐ multiple **AND**

Report options

☒ Add analysed data

☒ Add graph

☒ Add percentages Rows

0.05 significance level

OK Close

EXAMPLE (11.7) cont. (*sex-exam.pqs* file)

Hypotheses:

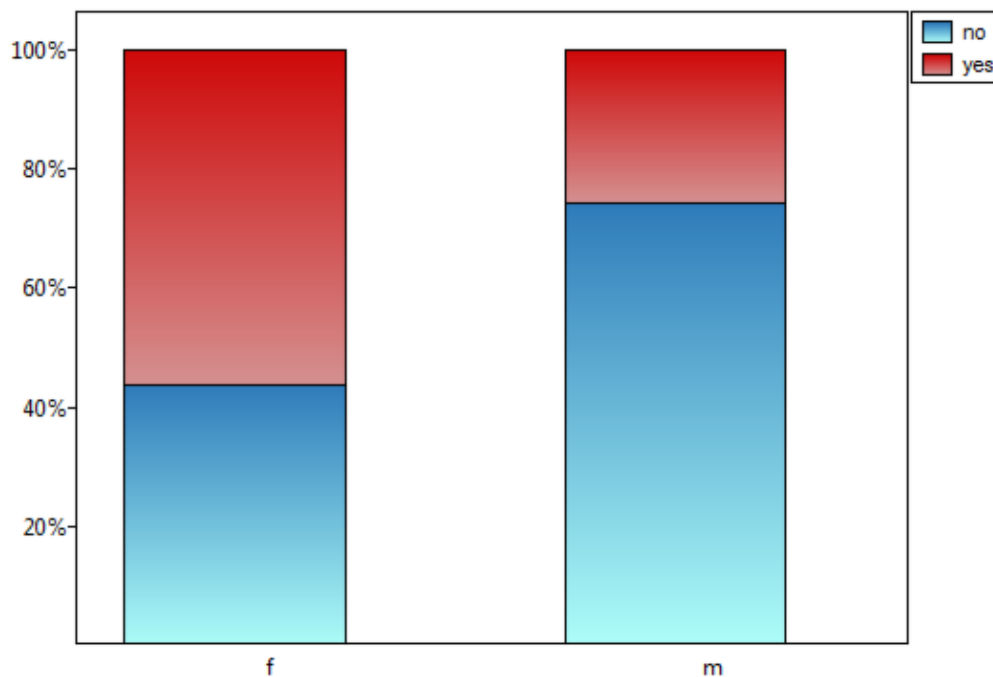
- \mathcal{H}_0 : there is no dependence between sex and exam passing in the analysed population,
 \mathcal{H}_1 : there is a dependence between sex and exam passing in the analysed population.

Fisher exact test, mid-p (2x2)	
Analysis time	0.02sec.
Analysed variables	sex;exam passed
Significance level	0.05
Size	170
Odds Ratio (exact)	0.268864
-95% CI for the Odds Ratio (exact)	0.130742
+95% CI for the Odds Ratio (exact)	0.537735
Odds Ratio (mid-p)	0.269805
-95% CI for the Odds Ratio (mid-p)	0.137445
+95% CI for the Odds Ratio (mid-p)	0.514591
Fisher	
One sided p-value	0.000043
Two sided p-value	0.000083
Mid-p	
2 * one sided p-value	0.000054

Expected:		
	no	yes
f	52.94	37.06
m	47.06	32.94

Data:		
	no	yes
f	40	50
m	60	20

Rows:		
	no	yes
f	44.44%	55.56%
m	75%	25%



The two-sided p value = 0.000083. So, using the Fisher exact test, similarly to the χ^2 test and the χ^2 test with the Yate's correction, on the significance level $\alpha = 0.05$ you accept the hypothesis informing us that there is a dependence between sex and exam passing in the analysed population. Significantly, the exam was passed more often by women ($\frac{50}{90} = 55.56\%$ out of all the women in the sample who passed the exam) than by men ($\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam).

The mid-p

The mid-p is the Fisher exact test correction. This modified p value is recommended by many statisticians (Lancaster 1961[48], Anscombe 1981[4], Pratt and Gibbons 1981[69], Plackett 1984[68], Miittinen 1985[63] and Barnard 1989[6], Rothman 2008[72]) as a method used in decreasing the Fisher exact test conservatism. As a result, using the mid-p the null hypothesis is rejected much more quickly than by using the Fisher exact test. For large samples a p value is calculated by using the χ^2 test with the Yate's correction and the Fisher test gives quite similar results. But a p value of the χ^2 test without any correction corresponds with the mid-p.

The p value of the mid-p is calculated by the transformation of the probability value for the Fisher exact test. The one-sided p value is calculated by using the following formula:

$$pI(\text{mid-p}) = pI(\text{Fisher}) - 0.5 \cdot P_{\text{point(given table)}},$$

where:

$pI(\text{mid-p})$ — one-sided p value of mid-p,

$pI(\text{Fisher})$ — one-sided p value of Fisher exact test,

and the two-sided p value is defined as a doubled value of the smaller one-sided probability:

$$pII(\text{mid-p}) = 2pI(\text{mid-p}),$$

where:

$pII(\text{mid-p})$ — two-sided p value of mid-p.

EXAMPLE (11.7) cont. (*sex-exam.pqs file*)

The two-sided p value of the contingency table from the (11.7) example is $p=0.000054$. So, on the significance level $\alpha=0.05$ (similarly to the Fisher exact test, the χ^2 test and χ^2 test with the Yate's correction) you accept the alternative hypothesis verifying that there is a dependence between sex and exam passing in the analysed population. Significantly, the exam was passed more often by women ($\frac{50}{90} = 55.56\%$ out of all the women in the sample who passed the exam) than by men ($\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam).

11.2.7 Relative Risk and Odds Ratio

The risk and odds designation of occurrence an analysed phenomenon, on the basis of exposure to the factor that can cause it, is estimated according to data collected in the contingency table 2×2 :

Table 11.4. The contingency table of 2×2 observed frequencies

Observed frequencies O_{ij}		Analysed phenomenon (illness)		
		occurs (case)	not occurs (control)	Total
Risk factor	exposed	O_{11}	O_{12}	$O_{11} + O_{12}$
	unexposed	O_{21}	O_{22}	$O_{21} + O_{22}$
	Total	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$n = O_{11} + O_{12} + O_{21} + O_{22}$

If a study is a **case-control** study, the **odds ratio** of occurrence the phenomenon is calculated for the table. Usually, they are retrospective studies – the researcher decides on his own about the sample size, with the phenomenon, and about the control sample (without the phenomenon).

If a study is a **cohort** study, the **relative risk** of occurrence the phenomenon is calculated for the table. Usually, they are prospective studies – the researcher cares about experiment conditions, because of the structure of an analysed phenomenon in a sample and in a population should be similar.

The odds ratio (2×2 table)

For the designation of odds ratio, we calculate the probability of being a case in the exposed group and in the unexposed group, according to the formulas:

$$odds_{exposed} = \frac{O_{11}/(O_{11} + O_{12})}{O_{12}/(O_{11} + O_{12})} = \frac{O_{11}}{O_{12}},$$

$$odds_{unexposed} = \frac{O_{21}/(O_{21} + O_{22})}{O_{22}/(O_{21} + O_{22})} = \frac{O_{21}}{O_{22}}.$$

The Odds Ratio:

$$OR = \frac{O_{11}/O_{12}}{O_{21}/O_{22}} = \frac{O_{11}O_{22}}{O_{12}O_{21}}.$$

The test of significance for the OR

This test is used to the hypothesis verification about the odds of occurrence the analysed phenomenon is the same in the group of exposed and unexposed to the risk factor.

Hypotheses:

$$\mathcal{H}_0 : OR = 1,$$

$$\mathcal{H}_1 : OR \neq 1.$$

The test statistic is defined by:

$$z = \frac{\ln(OR)}{SE},$$

where:

$$SE = \sqrt{\frac{1}{O_{11}} + \frac{1}{O_{12}} + \frac{1}{O_{21}} + \frac{1}{O_{22}}} - \text{standard error of the } \ln(OR).$$

The test statistic asymptotically (for large sample size) has the **normal distribution**.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

Note

In the interpretation of odds ratio significance, we usually use the designated confidence interval. Then, we check if the interval contains the value of 1.

The odds ratio, altogether with asymptotic confidence intervals, and the odds ratio significance test are calculated by:

- Chi-square test, OR/RR (2x2) window,
- Mantel-Heanszel OR/RR window – for each table designated by the strata.

Exact intervals and the **mid-p** intervals for the odds ratio are calculated by:

- Fisher exact test, mid-p (2x2) window.

The relative risk (2×2 table)

In the cohort study, we can designate the **risk** of occurrence the analysed phenomenon (because the structure of phenomenon, in the sample, should come closer to the population, from which the sample was taken) and calculate the relative risk (RR).

The estimated risk of occurrence the analysed phenomenon is designated by the following formula $R = \frac{O_{11}+O_{21}}{n}$. However, the relative risk is designated by:

$$RR = \frac{O_{11}/(O_{11} + O_{12})}{O_{21}/(O_{21} + O_{22})}$$

The test of significance for the RR

This test is used to the hypothesis verification about the risk of occurrence the analysed occurrence is the same in the group of exposed and unexposed to the risk factor.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : RR &= 1, \\ \mathcal{H}_1 : RR &\neq 1. \end{aligned}$$

The test statistic is defined by:

$$z = \frac{\ln(RR)}{SE},$$

where:

$$SE = \sqrt{\frac{1}{O_{11}} - \frac{1}{O_{21}+O_{22}} + \frac{1}{O_{21}} - \frac{1}{O_{21}+O_{22}}} - \text{standard error of the } \ln(RR).$$

The test statistic asymptotically (for large sample size) has the **normal distribution**.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

Note

In the interpretation of the relative risk significance, we usually use the designated confidence interval. Then, we check if the interval contains the value of 1.

The relative risk, altogether with the asymptotic confidence intervals, and the relative risk significance test are calculated by:

- Chi-square test, OR/RR (2x2) window,
- Mantel-Heanszel OR/RR window – for each table designated by the strata.

11.2.8 The Z test for 2 independent proportions

The Z test for 2 independent proportions is used in the similar situations as the *chi² test (2 × 2)*. It means, when there are 2 independent samples with the total size of n_1 and n_2 , with the 2 possible results to gain (one of the results is distinguished with the size of m_1 - in the first sample and m_2 - in the second one). For these samples it is also possible to calculate the distinguished proportions $p_1 = \frac{m_1}{n_1}$ and $p_2 = \frac{m_2}{n_2}$. This test is used to verify the hypothesis informing us that the distinguished proportions P_1 and P_2 in populations, from which the samples were drawn, are equal.

Basic assumptions:

- measurement on a **nominal scale** (alternatively: an **ordinal** or an **interval**),
- an **independent model**,
- large sample sizes.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : P_1 &= P_2, \\ \mathcal{H}_1 : P_1 &\neq P_2,\end{aligned}$$

where:

P_1, P_2 fraction for the first and the second population.

The test statistic is defined by:

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

where:

$$p = \frac{m_1 + m_2}{n_1 + n_2}.$$

The test statistic modified by the continuity correction is defined by:

$$Z = \frac{p_1 - p_2 - \frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

The Z Statistic with and without the continuity correction asymptotically (for the large sample sizes) has the **normal distribution**.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

Apart from the difference between proportions, the program calculates the value of the NNT.

NNT (*number needed to treat*) – indicator used in medicine to define the number of patients which have to be treated for a certain time in order to cure one person.

Note

From PQStat version 1.3.0, the confidence intervals for the difference between two independent proportions are estimated on the basis of the Newcombe-Wilson method. In the previous versions it was estimated on the basis of the Wald method.

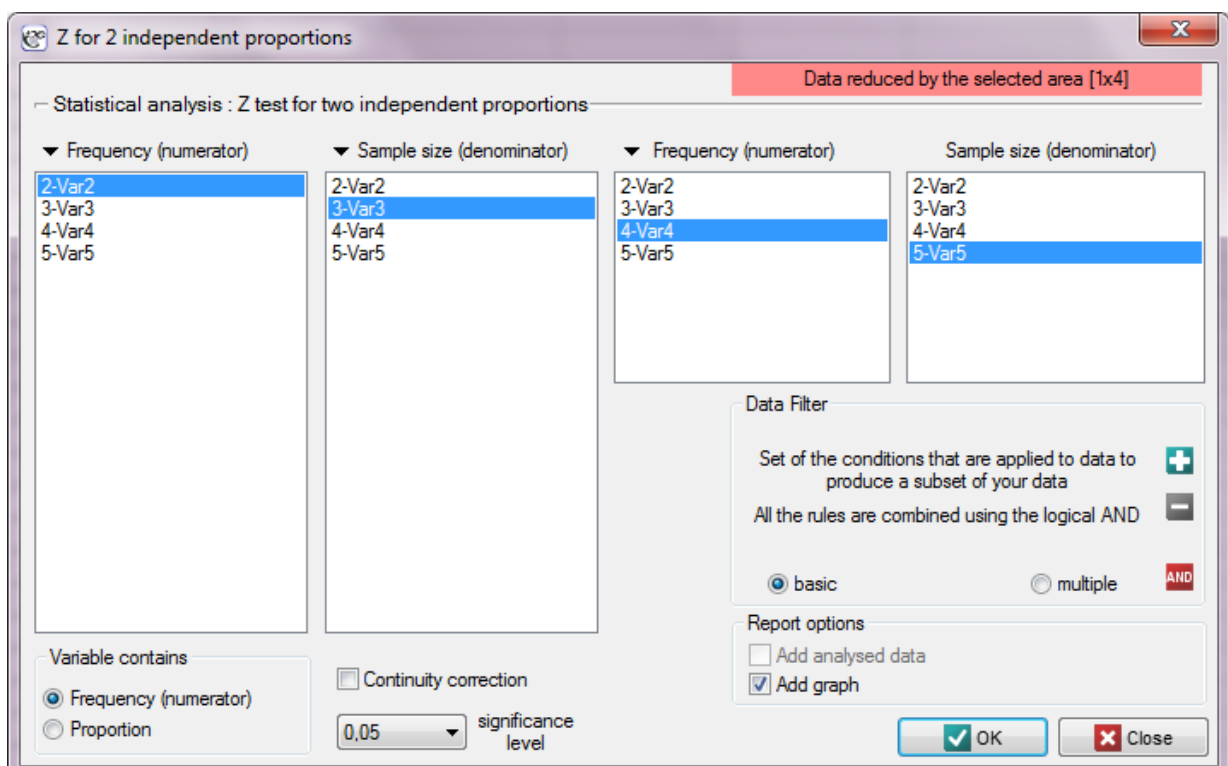
The justification of the change is as follows:

Confidence intervals based on the classical Wald method are suitable for large sample sizes and for the difference between proportions far from 0 or 1. For small samples and for the difference between proportions close to those extreme values, the Wald method can lead to unreliable results (Newcombe 1998[65], Miettinen 1985[64], Beal 1987[7], Wallenstein 1997[79]). A comparison and analysis of many methods which can be used instead of the simple Wald method can be found in Newcombe's study (1998)[65]. The suggested method, suitable also for extreme values of proportions, is the method first published by Wilson (1927)[86], extended to the intervals for the difference between two independent proportions.

Note

The confidence interval for the NNT is estimated on the basis of the Newcombe-Wilson method (Bender (2001)[8], Newcombe (1998)[65], Wilson (1927)[86]).

The settings window with the Z test for 2 proportions can be opened in Statistics menu \rightarrow NonParametric tests (ordered categories) $\rightarrow Z$ for 2 independent proportions.



EXAMPLE (11.7) cont. (sex-exam.pqs file)

You know that $\frac{50}{90} = 55.56\%$ out of all the women in the sample who passed the exam and $\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam. This data can be written in two ways — as a numerator and a denominator for each sample, or as a proportion and a denominator for each sample:

for Z test (frequency)	numerator - women	denominator - wome	numerator - men	denominator - men
	50	90	20	80
for Z test (proportion)	proportion - women	denominator - wome	proportion - men	denominator - men
	0.555555555556	90	0.25	80

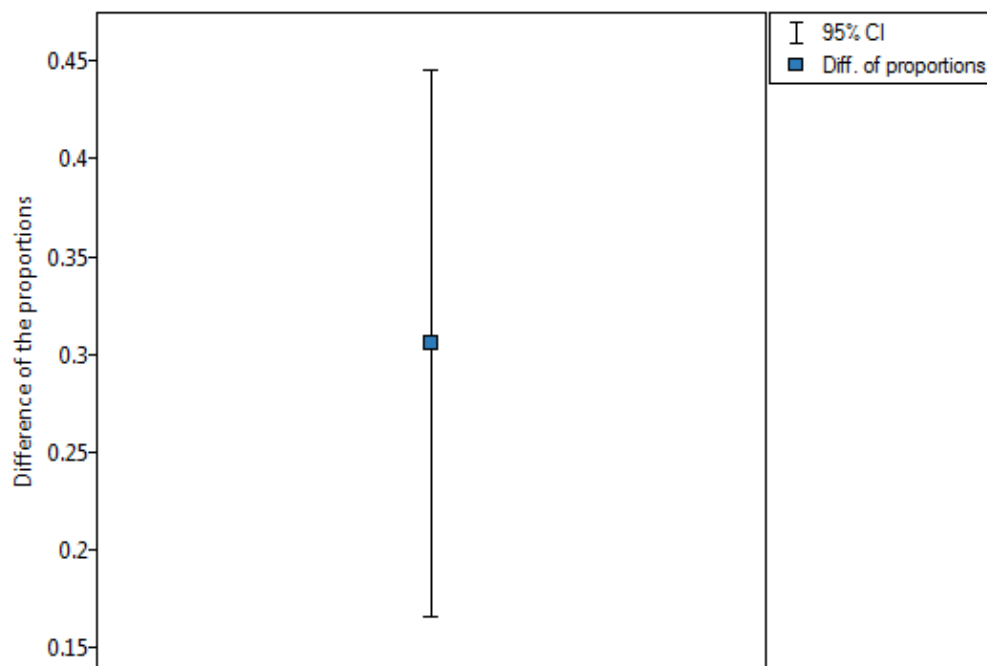
Hypotheses:

- \mathcal{H}_0 : The proportion of the men who passed the exam is the same as the proportion of the women who passed the exam in the analysed population,
- \mathcal{H}_1 : The proportion of the men who passed the exam is different than the proportion of the women who passed the exam in the analysed population.

Z test for two independent proportions	
Analysis time	0,08sec.
Analysed variables	Var2;Var3;Var4;Var5
Significance level	0,05
Continuity correction	No
Difference of the proportions	0,305556
-95% CI for the difference of the proportions	0,158695
+95% CI for the difference of the proportions	0,433518
NNT	3,272727
-95% CI NNT	2,306711
+95% CI NNT	6,301412
Z statistic	4,04047
p-value (asymptotic)	0,000053

Data:

v.1	v.2	v.3	v.4
50	90	20	80



Note

It is necessary to select the appropriate area (data without headings) before the analysis begins, because usually there are more information in a datasheet. You should also select the option indicating the content of the variable (frequency (numerator) or proportion). The difference between proportions distinguished in the sample is 30.56%, a 95% and the confidence interval for it (15.90%, 43.35%) does not contain 0.

Based on the Z test without the continuity correction as well as on the Z test with the continuity correction (p value = 0.000053 and p value = 0.0001), on the significance level $\alpha=0.05$, the alternative hypothesis can be accepted (similarly to the Fisher exact test, its the mid-p corrections, the χ^2 test and the χ^2 test with the Yate's correction). So, the proportion of men, who passed the exam is different than the proportion of women, who passed the exam in the analysed population. Significantly, the exam was passed more often by women ($\frac{50}{90} = 55.56\%$ out of all the women in the sample who passed the exam) than by men ($\frac{20}{80} = 25.00\%$ out of all the men in the sample who passed the exam).

EXAMPLE 11.8.

Let us assume that the mortality rate of a disease is 100% without treatment and that therapy lowers the mortality rate to 50% – that is the result of 20 years of study. We want to know how many people have to be treated to prevent 1 death in 20 years. To answer that question, two samples of 100 people were taken from the population of the diseased. In the sample without treatment there are 100 patients of whom we know they will all die without the therapy. In the sample with therapy we also have 100 patients of whom 50 will survive.

Patients – not undergoing therapy		Patients – undergoing therapy	
sample numerator	sample (denominator)	sample numerator	sample (denominator)
100	100	50	100

We will calculate the NNT.

Z test for two independent proportions	
Analysis time	0,04sec.
Analysed variables	Var2;Var3;Var4;Var5
Significance level	0,05
Continuity correction	Yes
Difference of the proportions	0,5
-95% CI for the difference of the proportions	0,396962
+95% CI for the difference of the proportions	0,596168
NNT	2
-95% CI NNT	1,677378
+95% CI NNT	2,519135
Z statistic	8,001666
p-value (asymptotic)	<0.000001

Data:

v.1	v.2	v.3	v.4
100	100	50	100

The difference between proportions is statistically significant ($p < 0.000001$) but we are interested in the NNT – its value is 2, so the treatment of 2 patients for 20 years will prevent 1 death. The calculated confidence interval value of 95% should be rounded off to a whole number, wherefore the NNT is 2 to 3 patients.

11.2.9 The McNemar test, the Bowker test of internal symmetry

Basic assumptions:

- measurement on a **nominal scale**,

- a **dependent model**.

The McNemar test

The McNemar test (McNemar (1947)[61]) is used to verify the hypothesis determining the agreement between the results of the measurements, which were done twice $X^{(1)}$ and $X^{(2)}$ of an X feature (between 2 dependent variables $X^{(1)}$ and $X^{(2)}$). The analysed feature can have only 2 categories (defined here as **(+)** and **(-)**). The McNemar test can be calculated on the basis of **raw data** or on the basis of a 2×2 **contingency table**.

Table 11.5. 2×2 contingency table for the observed frequencies of dependent variables

Observed frequencies O_{ij}		$X^{(2)}$		
		(+)	(-)	Total
$X^{(1)}$	(+)	O_{11}	O_{12}	$O_{11} + O_{12}$
	(-)	O_{21}	O_{22}	$O_{21} + O_{22}$
	Total	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$n = O_{11} + O_{12} + O_{21} + O_{22}$

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : O_{12} &= O_{21}, \\ \mathcal{H}_1 : O_{12} &\neq O_{21}.\end{aligned}$$

The test statistic is defined by:

$$\chi^2 = \frac{(O_{12} - O_{21})^2}{O_{12} + O_{21}}.$$

This statistic asymptotically (for large frequencies) has the χ^2 **distribution** with a 1 degree of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned}\text{if } p &\leq \alpha \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p &> \alpha \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The Continuity correction for the McNemar test

This correction is a more conservative test than the McNemar test (a null hypothesis is rejected much more rarely than when using the McNemar test). It guarantees the possibility of taking in all the values of real numbers by the test statistic, according to the χ^2 distribution assumption. Some sources give the information that the continuity correction should be used always, but some other ones inform, that only if the frequencies in the table are small.

The test statistic with the continuity correction is defined by:

$$\chi^2 = \frac{(|O_{12} - O_{21}| - 1)^2}{O_{12} + O_{21}}.$$

Odds ratio of a result change

If the study is carried out twice for the same feature and on the same objects – then, **odds ratio** for the result change (from **(+)** to **(-)** and inversely) is calculated for the table.

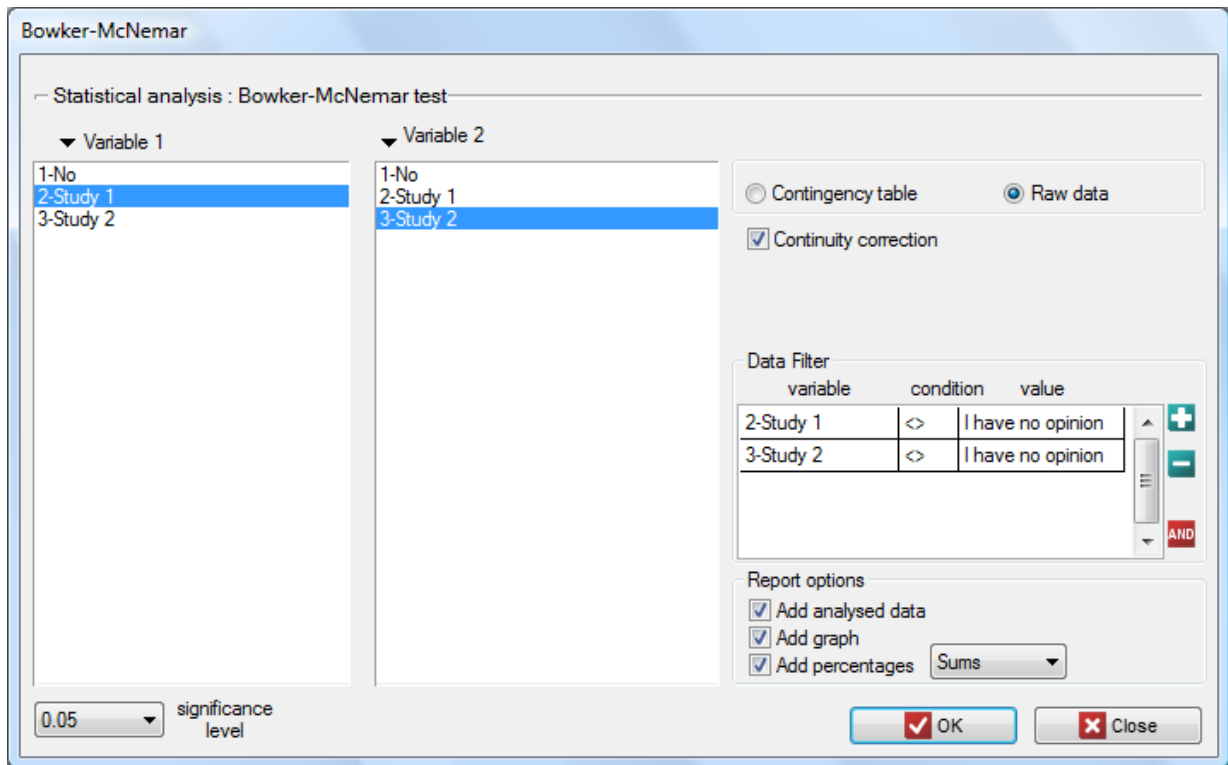
The odds for the result change from **(+)** to **(-)** is O_{12} , and the odds for the result change from **(-)** to **(+)** is O_{21} . Odds Ratio (OR) is:

$$OR = \frac{O_{12}}{O_{21}}.$$

Confidence interval for the odds ratio is calculated on the base of the standard error:

$$SE = \sqrt{\frac{1}{O_{12}} + \frac{1}{O_{21}}}.$$

The settings window with the Bowker-McNemar test can be opened in Statistics menu → NonParametric tests (unordered categories) → Bowker-McNemar or in [Wizard](#).



The Bowker test of internal symmetry

The Bowker test of internal symmetry (Bowker (1948)[11]) is an extension of the McNemar test for 2 variables with more than 2 categories ($c > 2$). It is used to verify the hypothesis determining the symmetry of 2 results of measurements executed twice $X^{(1)}$ and $X^{(2)}$ of X feature (symmetry of 2 dependent variables $X^{(1)}$ i $X^{(2)}$). An analysed feature may have more than 2 categories. The Bowker test of internal symmetry can be calculated on the basis of either [raw data](#) or a $c \times c$ [contingency table](#).

Table 11.6. $c \times c$ contingency table for the observed frequencies of dependent variables

Observed frequencies O_{ij}		$X^{(2)}$				
		$X_1^{(2)}$	$X_2^{(2)}$...	$X_c^{(2)}$	Total
$X^{(1)}$	$X_1^{(1)}$	O_{11}	O_{12}	...	O_{1c}	$\sum_{j=1}^c O_{1j}$
	$X_2^{(1)}$	O_{21}	O_{22}	...	O_{2c}	$\sum_{j=1}^c O_{2j}$

	$X_c^{(1)}$	O_{c1}	O_{c2}	...	O_{cc}	$\sum_{j=1}^c O_{cj}$
	Total	$\sum_{i=1}^c O_{i1}$	$\sum_{i=1}^c O_{i2}$...	$\sum_{i=1}^c O_{ic}$	$n = \sum_{i=1}^c \sum_{j=1}^c O_{ij}$

Hypotheses:

$$\mathcal{H}_0 : O_{ij} = O_{ji},$$

$$\mathcal{H}_1 : O_{ij} \neq O_{ji} \text{ for at least one pair } O_{ij}, O_{ji},$$

where $j \neq i, j \in 1, 2, \dots, c, i \in 1, 2, \dots, c$, so O_{ij} and O_{ji} are the frequencies of the symmetrical pairs in the $c \times c$ table

The test statistic is defined by:

$$\chi^2 = \sum_{i=1}^c \sum_{j>i} \frac{(O_{ij} - O_{ji})^2}{O_{ij} + O_{ji}}.$$

This statistic asymptotically (for large sample size) has the χ^2 distribution with a number of degrees of freedom calculated using the formula: $df = \frac{c(c-1)}{2}$.

The p value, designated on the basis of the test statistic, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

EXAMPLE 11.9. (opinion.pqs file)

Two different surveys were carried out. They were supposed to analyse students' opinions about the particular academic professor. Both the surveys enabled students to give a positive opinion, a negative and a neutral one. Both surveys were carried out on the basis of the same sample of 250 students. But the first one was carried out the day before an exam done by the professor, and the other survey the day after the exam. There are some data below – in a form of raw rows, and all the data – in the form of a contingency table. Check, if both surveys give the similar results.

	Study 1	Study 2
I have no opinion		positive
positive		negative
positive		negative
positive		positive

	negative	positive	I have no opinion
negative	50	4	3
positive	44	54	5
I have no opinion	35	18	37

Hypotheses:

\mathcal{H}_0 : the number of students, who changed their opinions is exactly the same for each of the possible symmetric opinion changes,

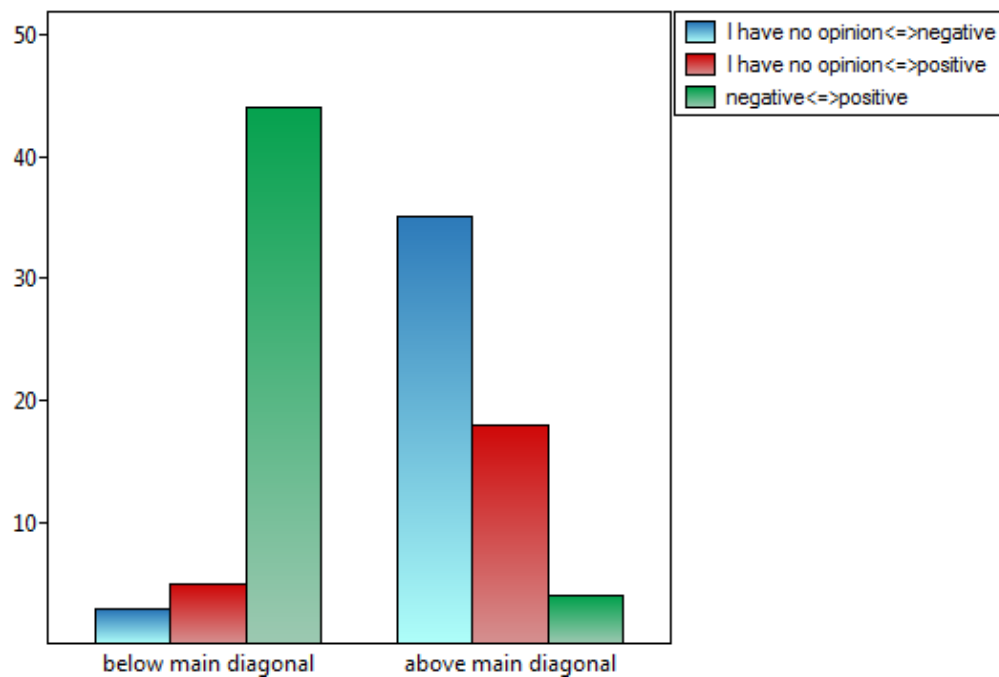
\mathcal{H}_1 : the number of students, who changed their opinions is different for at least one of the possible symmetric opinion changes,

where, for example, changing the opinion from positive to negative one is symmetrical to changing the opinion from negative to positive one.

Bowker-McNemar test	
Analysis time	0.03sec.
Analysed variables	Study 1;Study 2
Significance level	0.05
Continuity correction	Yes
Size = number of pairs	250
Chi-square statistic	63.2378432
Degrees of freedom	3
p-value	<0.0000001

Data:	I have no	negative	positive
I have no	37	35	18
negative	3	50	4
positive	5	44	54

Sums:	I have no	negative	positive
I have no	14.8%	14%	7.2%
negative	1.2%	20%	1.6%
positive	2%	17.6%	21.6%



Comparing the p value for the Bowker test ($p \text{ value} < 0.000001$) with the significance level $\alpha = 0.05$ it may be assumed that students changed their opinions. Looking at the table you can see that, there were more students who changed their opinions to negative ones after the exam, than those who changed it to positive ones after the exam. There were also students who did not evaluate the professor in the positive way after the exam any more.

If you limit your analysis only to the people having clear opinions about the professor (positive or negative ones), you can use the McNemar test:

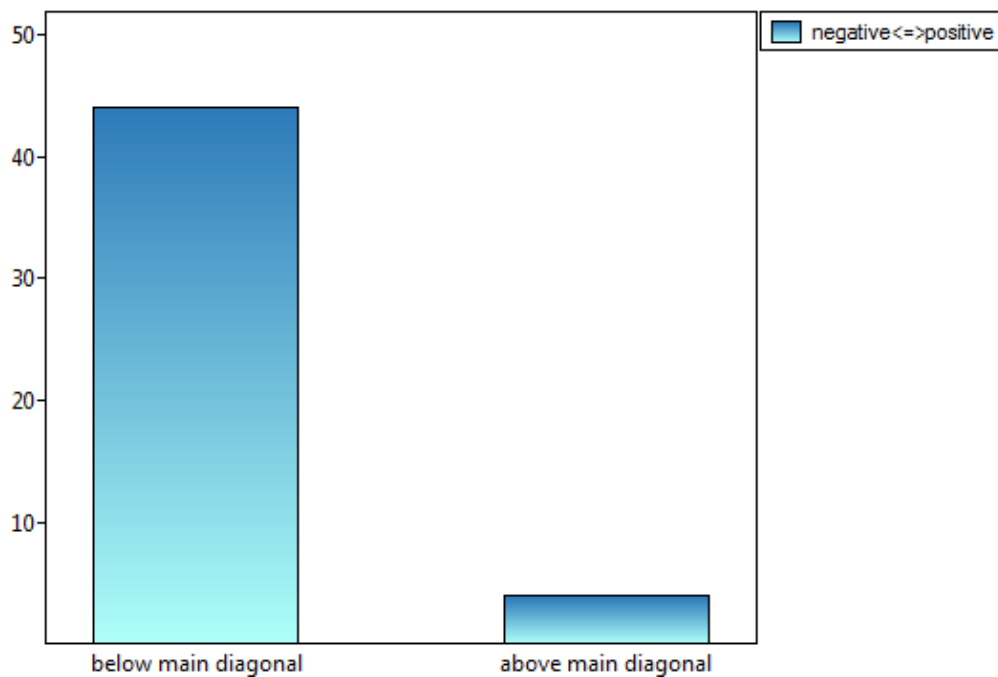
Hypotheses:

- \mathcal{H}_0 : the number of students, who changed their opinions from negative to positive ones is exactly the same as those, who changed their opinions from positive to negative,
- \mathcal{H}_1 : the number of students, who changed their opinions from negative to positive ones is different from those, who changed their opinions from positive to negative.

Bowker-McNemar test	
Analysis time	0.04sec.
Analysed variables	Study 1;Study 2
Significance level	0.05
Continuity correction	Yes
Data Filter	Study 1<>I have no opinion and Study 2<>I have no opi
Size = number of pairs	152
Iloraz szans	0.090909
-95% CI for the Odds Ratio	0.032665
+95% CI for the Odds Ratio	0.253007
Chi-square statistic	31.6875
Degrees of freedom	1
p-value	<0.000001

Data:		
	negative	positive
negative	50	4
positive	44	54

Sums:		
	negative	positive
negative	32.89%	2.63%
positive	28.95%	35.53%



If you compare the p value, calculated for the McNemar test (p value < 0.000001), with the significance level $\alpha = 0.05$, you draw the conclusion that the students changed their opinions. There were much more students, who changed their opinions to negative ones after the exam, than those who changed their opinions to positive ones. The possibility of changing the opinion from positive (before the exam) to negative (after the exam) is eleven ($\frac{44}{4}$) times greater than from negative to positive (the chance to change opinion in the opposite direction is: $(\frac{4}{44}) = 0.090909$).

11.2.10 Z Test for two dependent proportions

Z Test for two dependent proportions is used in situations similar to the **McNemar's Test**, i.e. when we have 2 dependent groups of measurements ($X^{(1)}$ i $X^{(2)}$), in which we can obtain 2 possible results of the studied feature (**(+)**(**(-)**).

Observed sizes O_{ij}		$X^{(2)}$		
		(+)	(-)	Suma
$X^{(1)}$	(+)	O_{11}	O_{12}	$O_{11} + O_{12}$
	(-)	O_{21}	O_{22}	$O_{21} + O_{22}$
	Sum	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$n = O_{11} + O_{12} + O_{21} + O_{22}$

We can also calculated distinguished proportions for those groups $p_1 = \frac{O_{11}+O_{12}}{n}$ i $p_2 = \frac{O_{11}+O_{21}}{n}$. The test serves the purpose of verifying the hypothesis that the distinguished proportions P_1 and P_2 in the population from which the sample was drawn are equal.

Basic assumptions:

- measurement on the **nominal**, **ordinal**, or **interval scale**,
- **dependent model**,
- large sample size.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : P_1 - P_2 &= 0, \\ \mathcal{H}_1 : P_1 - P_2 &\neq 0,\end{aligned}$$

where:

P_1, P_2 fractions for the first and the second measurement.

The test statistic has the form presented below:

$$Z = \frac{p_1 - p_2}{\sqrt{O_{21} + O_{12}}} \cdot n,$$

The Z Statistic asymptotically (for the large sample size) has the [normal distribution](#).

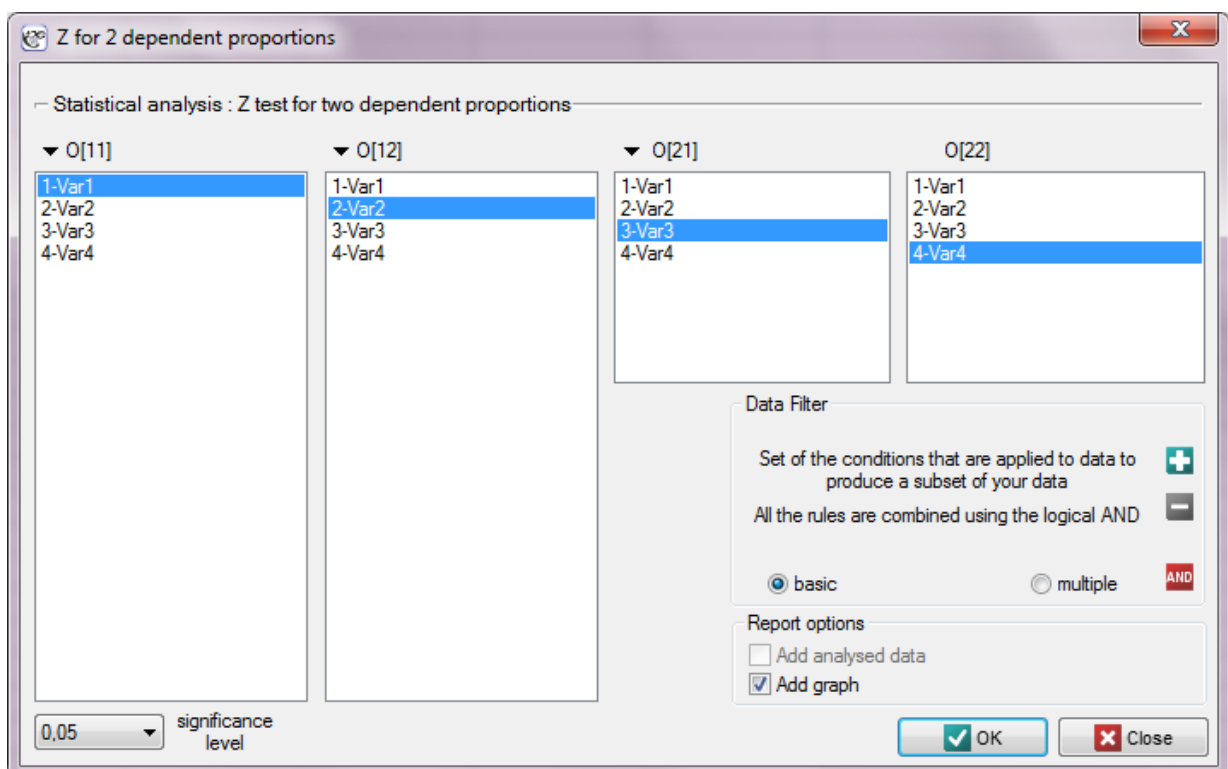
On the basis of [test statistics](#), [p value](#) is estimated and then compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

Note

Confidence interval for the difference of two dependent proportions is estimated on the basis of the Newcombe-Wilson method.

The window with settings for Z-Test for two dependent proportions is accessed via the menu Statistics→Nonparametric tests (nonordered categories)→Z-Test for two dependent proportions.



EXAMPLE (11.9) cont. (file *opinia.pqs*)

When we limit the study to people who have a specific opinion about the professor (i.e. those who

only have a positive or a negative opinion) we will have 152 such students. The data for calculations are: $O_{11} = 50$, $O_{12} = 4$, $O_{21} = 44$, $O_{22} = 54$. We know that $\frac{50+4}{152} = 35.53\%$ students expressed a negative opinion before the exam. After the exam the percentage was $\frac{50+44}{152} = 61.84\%$.

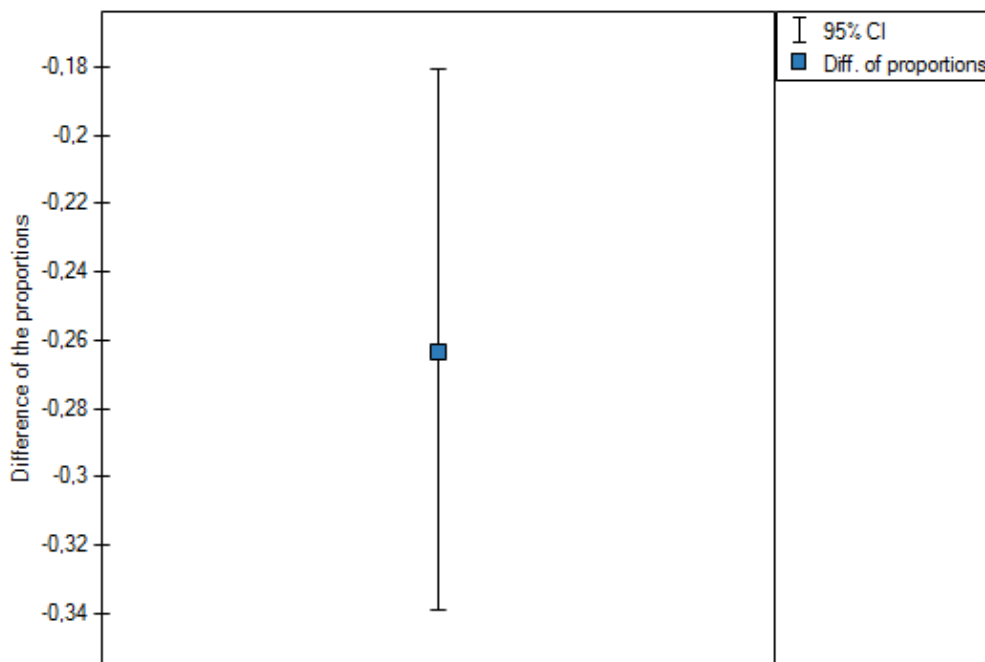
Hypotheses:

- \mathcal{H}_0 : a lack of a difference between the number of negative evaluations of the professor before and after the exam,
 \mathcal{H}_1 : there is a difference between the number of negative evaluations of the professor before and after the exam.

Z test for two dependent proportions	
Analysis time	0,18sec.
Analysed variables	Var1;Var2;Var3;Var4
Significance level	0,05
Continuity correction	No
O[11]+O[12]	54
O[11]+O[21]	94
Proportion 1	0,355263
Proportion 2	0,618421
Difference of the proportions	-0,263158
-95% CI for the difference of the proportions	-0,338845
+95% CI for the difference of the proportions	-0,180717
Z statistic	-5,773503
p-value (asymptotic)	<0.000001

Data:

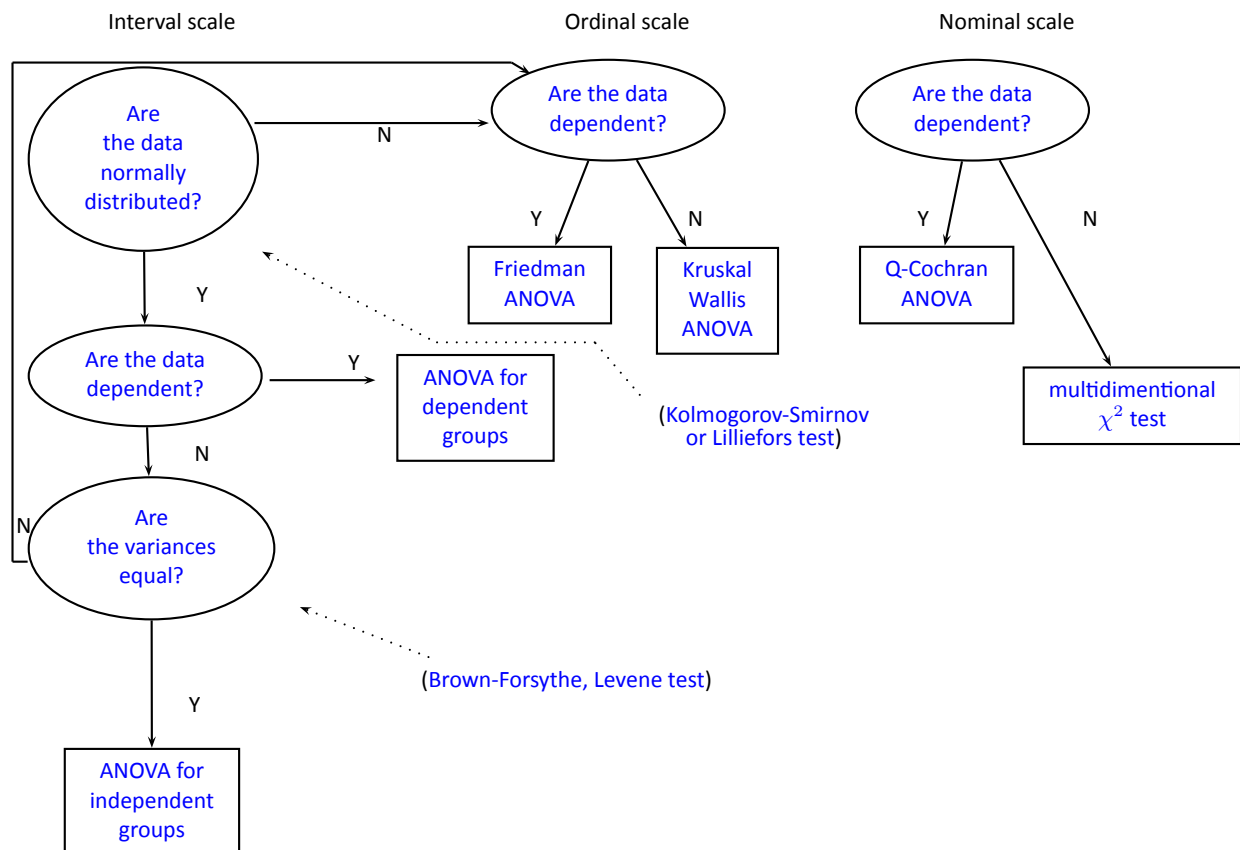
v.1	v.2	v.3	v.4
50	4	44	54



The difference in proportions distinguished in the sample is 26.32%, and the confidence interval of 95% for the sample (18.07%, 33.88%) does not contain 0.

On the basis of a Z test ($p=0.0001$), on the significance level of $\alpha=0.05$ (similarly to the case of McNemar's test) we accept the alternative hypothesis. Therefore, the proportion of negative evaluations before the exam differs from the proportion of negative evaluations after the exam. Indeed, after the exam there are more negative evaluations of the professor.

12 COMPARISON - MORE THAN 2 GROUPS



Note

Note, that simultaneous comparison of more than two groups can NOT be replaced with multiple performance the [tests for the comparison of two groups](#). It is the result of the necessity of controlling the [I type error \$\alpha\$](#) . Choosing the α and using the k -fold selected test for the comparison of 2 groups, we could make the assumed level much higher α . It is possible to avoid this error using the ANOVA test (Analysis of Variance) and contrasts or the POST-HOC tests dedicated to them.

12.1 PARAMETRIC TESTS

12.1.1 The ANOVA for independent groups

The one-way analysis of variance (ANOVA for independent groups) proposed by Ronald Fisher, is used to verify the hypothesis determining the equality of **means** of an analysed variable in several ($k \geq 2$) populations.

Basic assumptions:

- measurement on an **interval scale**,
- **normality of distribution** of an analysed feature in each population,
- an **independent model**,
- **equality of variances** of an analysed variable in all populations.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \mu_1 = \mu_2 = \dots = \mu_k, \\ \mathcal{H}_1 : & \text{not all } \mu_j \text{ are equal } (j = 1, 2, \dots, k),\end{aligned}$$

where:

$\mu_1, \mu_2, \dots, \mu_k$ – means of an analysed variable of each population.

The test statistic is defined by:

$$F = \frac{MS_{BG}}{MS_{WG}},$$

where:

$$MS_{BG} = \frac{SS_{BG}}{df_{BG}} - \text{mean square between-groups},$$

$$MS_{WG} = \frac{SS_{WG}}{df_{WG}} - \text{mean square within-groups},$$

$$SS_{BG} = \sum_{j=1}^k \frac{(\sum_{i=1}^{n_j} x_{ij})^2}{n_j} - \frac{(\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij})^2}{N} - \text{between-groups sum of squares},$$

$$SS_{WG} = SS_T - SS_{BG} - \text{within-groups sum of squares},$$

$$SS_T = \left(\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}^2 \right) - \frac{(\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij})^2}{N} - \text{total sum of squares},$$

$$df_{BG} = k - 1 - \text{between-groups degrees of freedom},$$

$$df_{WG} = df_T - df_{BG} - \text{within-groups degrees of freedom},$$

$$df_T = N - 1 - \text{total degrees of freedom},$$

$$N = \sum_{j=1}^k n_j,$$

$$n_j - \text{samples sizes for } (j = 1, 2, \dots, k),$$

$$x_{ij} - \text{values of a variable taken from a sample for } (i = 1, 2, \dots, n_j), (j = 1, 2, \dots, k).$$

The F statistic has the **F Snedecor distribution** with df_{BG} and df_{WG} degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha & \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

12.1.2 The contrasts and the POST-HOC tests

An analysis of the variance enables you to get information only if there are any significant differences among populations. It does not inform you which populations are different from each other. To gain some more detailed knowledge about the differences in particular parts of our complex structure, you should use **contrasts** (if you do the earlier planned and usually only particular comparisons), or the procedures of multiple comparisons **POST-HOC tests** (when having done the analysis of variance, we look for differences, usually between all the pairs).

The number of all the possible simple comparisons is calculated using the following formula:

$$c = \binom{k}{2} = \frac{k(k-1)}{2}$$

Hypotheses:

The first example - **simple comparisons** (comparison of 2 selected means):

$$\begin{aligned}\mathcal{H}_0 : \mu_1 &= \mu_2, \\ \mathcal{H}_1 : \mu_1 &\neq \mu_2.\end{aligned}$$

The second example - **complex comparisons** (comparison of combination of selected means):

$$\begin{aligned}\mathcal{H}_0 : \mu_1 &= \frac{\mu_2 + \mu_3}{2}, \\ \mathcal{H}_1 : \mu_1 &\neq \frac{\mu_2 + \mu_3}{2}.\end{aligned}$$

If you want to define the selected hypothesis you should ascribe the contrast value c_j , ($j = 1, 2, \dots, k$) to each mean. The c_j values are selected, so that their sums of compared sides are the opposite numbers, and their values of means which are not analysed count to 0.

The first example: $c_1 = 1, c_2 = -1, c_3 = 0, \dots, c_k = 0$.

The second example: $c_1 = 2, c_2 = -1, c_3 = -1, c_4 = 0, \dots, c_k = 0$.

How to choose the proper hypothesis:

- (i) Comparing the differences between the selected means with the **critical difference (CD)** calculated using the proper POST-HOC test:

$$\begin{aligned}\text{if the differences between means} &\geq CD \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if the differences between means} &< CD \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

- (ii) Comparing the **p value**, designated on the basis of the **test statistic** of the proper POST-HOC test, with the significance level α :

$$\begin{aligned}\text{if } p &\leq \alpha \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p &> \alpha \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The LSD Fisher test

For simple and complex comparisons, equal-size groups as well as unequal-size groups.

- (i) The value of critical difference is calculated by using the following formula:

$$CD = \sqrt{F_{\alpha, 1, df_{WG}}} \cdot \sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n_j} \right) MS_{WG}},$$

where:

$F_{\alpha,1,df_{WG}}$ - is the **critical value** (statistic) of the **F Snedecor distribution** for a given significance level α and degrees of freedom, adequately: 1 and df_{WG} .

(ii) The test statistic is defined by:

$$t = \frac{\sum_{j=1}^k c_j \bar{x}_j}{\sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n_j}\right) MS_{WG}}}.$$

The test statistic has the **t-Student distribution** with df_{WG} degrees of freedom.

The Scheffe test

For simple comparisons, equal-size groups as well as unequal-size groups.

(i) The value of a critical difference is calculated by using the following formula:

$$CD = \sqrt{F_{\alpha,df_{BG},df_{WG}}} \cdot \sqrt{(k-1) \left(\sum_{j=1}^k \frac{c_j^2}{n_j}\right) MS_{WG}},$$

where:

$F_{\alpha,df_{BG},df_{WG}}$ - is the **critical value** (statistic) of the **F Snedecor distribution** for a given significance level α and df_{BG} and df_{WG} degrees of freedom.

(ii) The test statistic is defined by:

$$F = \frac{\left(\sum_{j=1}^k c_j \bar{x}_j\right)^2}{(k-1) \left(\sum_{j=1}^k \frac{c_j^2}{n_j}\right) MS_{WG}}.$$

The test statistic has the **F Snedecor distribution** with df_{BG} and df_{WG} degrees of freedom.

The Tukey test.

For simple comparisons, equal-size groups as well as unequal-size groups.

(i) The value of a critical difference is calculated by using the following formula:

$$CD = \frac{\sqrt{2} \cdot q_{\alpha,df_{WG},k} \cdot \sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n_j}\right) MS_{WG}}}{2},$$

where:

$q_{\alpha,df_{WG},k}$ - is the **critical value** (statistic) of the studentized range distribution for a given significance level α and df_{WG} and k degrees of freedom.

(ii) The test statistic is defined by:

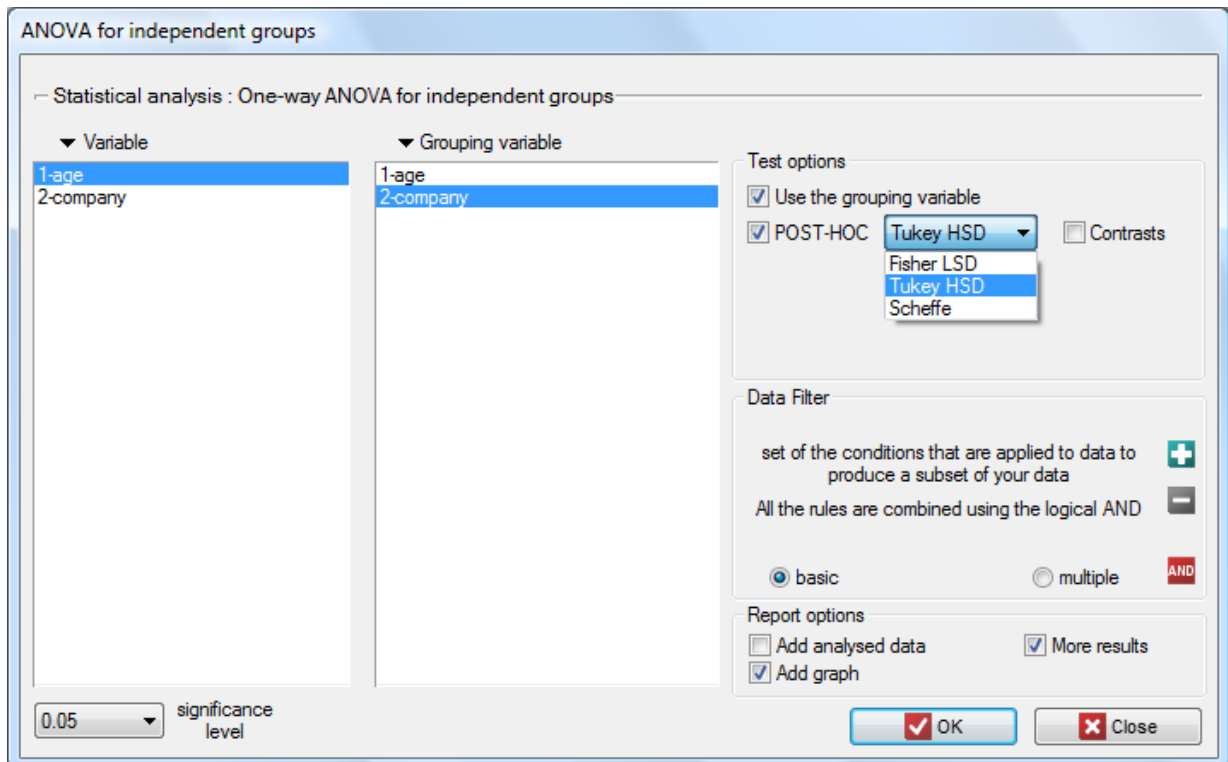
$$q = \sqrt{2} \frac{\sum_{j=1}^k c_j \bar{x}_j}{\sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n_j}\right) MS_{WG}}}.$$

The test statistic has the studentized range distribution with df_{WG} and k degrees of freedom.

Info.

The algorithm for calculating the p value and the statistic of the studentized range distribution in PQStat is based on the Lund works (1983)[54]. Other applications or web pages may calculate a little bit different values than PQStat, because they may be based on less precised or more restrictive algorithms (Copenhaver and Holland (1988), Gleason (1999)).

The settings window with the One-way ANOVA for independent groups can be opened in Statistics menu→Parametric tests→ANOVA for independent groups or in [Wizard](#).

**EXAMPLE 12.1.** (age ANOVA.pqs file)

There are 150 persons chosen randomly from the population of workers of 3 different transport companies. From each company there are 50 persons drawn to the sample. Before the experiment begins, you should check if the average age of the workers of these companies is similar, because the next step of the experiment depends on it. The age of each participant is written in years.

Age (company 1): 27, 33, 25, 32, 34, 38, 31, 34, 20, 30, 30, 27, 34, 32, 33, 25, 40, 35, 29, 20, 18, 28, 26, 22, 24, 24, 25, 28, 32, 32, 33, 32, 34, 27, 34, 27, 35, 28, 35, 34, 28, 29, 38, 26, 36, 31, 25, 35, 41, 37

Age (company 2): 38, 34, 33, 27, 36, 20, 37, 40, 27, 26, 40, 44, 36, 32, 26, 34, 27, 31, 36, 36, 25, 40, 27, 30, 36, 29, 32, 41, 49, 24, 36, 38, 18, 33, 30, 28, 27, 26, 42, 34, 24, 32, 36, 30, 37, 34, 33, 30, 44, 29


Age (company 3): 34, 36, 31, 37, 45, 39, 36, 34, 39, 27, 35, 33, 36, 28, 38, 25, 29, 26, 45, 28, 27, 32, 33, 30, 39, 40, 36, 33, 28, 32, 36, 39, 32, 39, 37, 35, 44, 34, 21, 42, 40, 32, 30, 23, 32, 34, 27, 39, 37, 35

Before you do this example, it is worth starting with the similar task but related to 2 groups only ([11.7](#)).

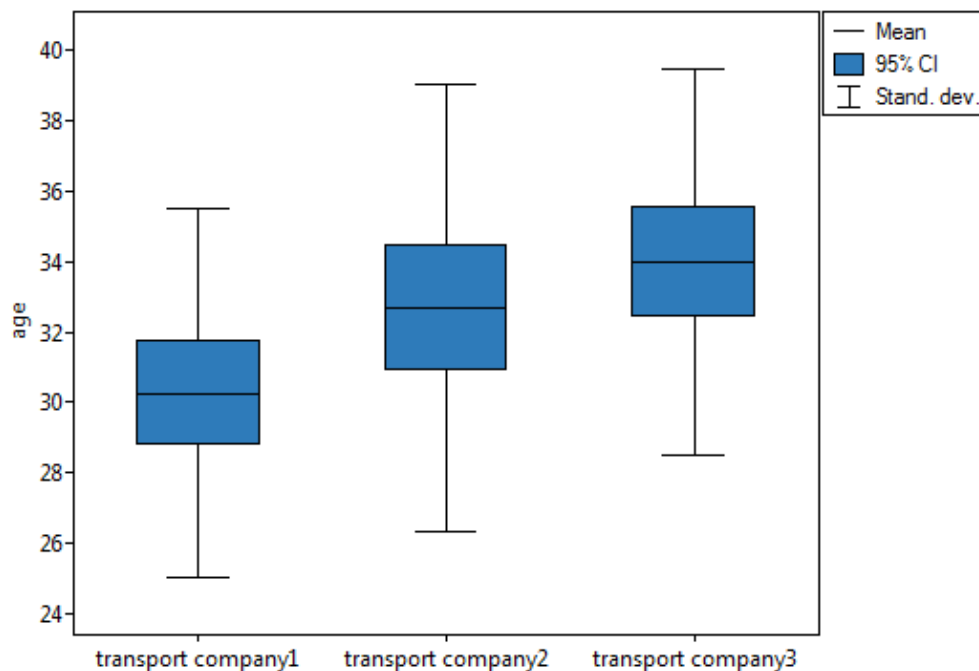
Hypotheses:

- \mathcal{H}_0 : the average age of the workers off all the analysed transport companies is the same,
- \mathcal{H}_1 : at least 2 means are different.

One-way ANOVA for independent groups	
Analysis time	0.11sec.
Analysed variables	age,company
Significance level	0.05
Grouping variable	company(transport comp
Group name	transport company1
Group size	50
Group mean	30.26
Group standard deviation	5.23259
Std. err. of the group mean	0.74
-95% CI for the group mean	28.772914
+95% CI for the group mean	31.747086
Group name	transport company2
Group size	50
Group mean	32.68
Group standard deviation	6.358154
Std. err. of the group mean	0.899179
-95% CI for the group mean	30.873033
+95% CI for the group mean	34.486967
Group name	transport company3
Group size	50
Group mean	33.98
Group standard deviation	5.482775
Std. err. of the group mean	0.775381
-95% CI for the group mean	32.421813
+95% CI for the group mean	35.538187
Total sum of squares (SS[T])	5151.893333
Between-groups sum of squares (SS[BG])	356.413333
Within-groups sum of squares (SS[WG])	4795.48
Mean square between-groups (MS[BG])	178.206667
Mean square within-groups (MS[WG])	32.622313
Between-groups degrees of freedom (df[BG])	2
Within-groups degrees of freedom (df[WG])	147
Total degrees of freedom (df[T])	149
F statistic	5.462723
p-value	0.005147

Comparing the p value = 0.005147 of the one-way analysis of variance with the significance level $\alpha = 0.05$, you can draw the conclusion that the average ages of workers of these transport companies is not the same. Based just on the ANOVA result, you do not know precisely which groups differ from others in terms of age. To gain such knowledge, it must be used one of the POST-HOC tests, for example the [Tukey test](#). To do this, you should resume the analysis by clicking  and then, in the options window for the test, you should select Tukey HSD and Add graph.

POST-HOC (Tukey HSD)			
	transport company1	transport company2	transport company3
Difference of the means			
transport company1		2.42	3.72
transport company2	2.42		1.3
transport company3	3.72	1.3	
CD			
transport company1		2.73086	2.73086
transport company2	2.73086		2.73086
transport company3	2.73086	2.73086	
Statistic q			
transport company1		2.99601	4.60543
transport company2	2.99601		1.60943
transport company3	4.60543	1.60943	
p-value			
transport company1		0.08965	0.00403
transport company2	0.08965		0.49229
transport company3	0.00403	0.49229	



The critical difference (CD) calculated for each pair of comparisons is the same (because the groups sizes are equal) and counts to 2.730855. The comparison of the CD value with the value of the mean difference indicates, that there are significant differences only between the mean age of the workers from the first and the third transport company (only if these 2 groups are compared, the CD value is less than the difference of the means). The same conclusion you draw, if you compare the p value of POST-HOC test with the significance level $\alpha = 0.05$. The workers of the first transport company are about 3 years younger (on average) than the workers of the third transport company.

Note

The assumptions for the single-factor analysis of variance are fulfilled:

- the age has the normal distribution in each of the analysed transport company (the p value of the **Lilliefors test** adequately counts to: $p = 0.134516$, $p = 0.603209$ and $p = 0.607648$),
- the **Brown-Forsythe test** indicates that there are no significant differences in the variances of the transport companies workers' age ($p = 0.430173$).

12.1.3 The Brown-Forsythe test and the Levene test

Both tests: the **Levene test** (Levene, 1960 [50]) and the **Brown-Forsythe test** (Brown and Forsythe, 1974 [16]) are used to verify the hypothesis determining the equality of **variance** of an analysed variable in several ($k \geq 2$) populations.

Basic assumptions:

- measurement on an **interval scale**,
- **normality of distribution** of an analysed feature in each population,
- an **independent model**.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2, \\ \mathcal{H}_1 : & \text{not all } \sigma_j^2 \text{ are equal } (j = 1, 2, \dots, k),\end{aligned}$$

where:

$\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$ – variances of an analysed variable of each population.

The analysis is based on calculating the absolute deviation of measurement results from the mean (in the Levene test) or from the median (in the Brown-Forsythe test), in each of the analysed groups. This absolute deviation is the set of data which are under the same procedure performed to the **analysis of variance for independent groups**. Hence, the test statistic is defined by:

$$F = \frac{MS_{BG}}{MS_{WG}},$$

The test statistic has the **F Snedecor distribution** with df_{BG} and df_{WG} degrees of freedom.

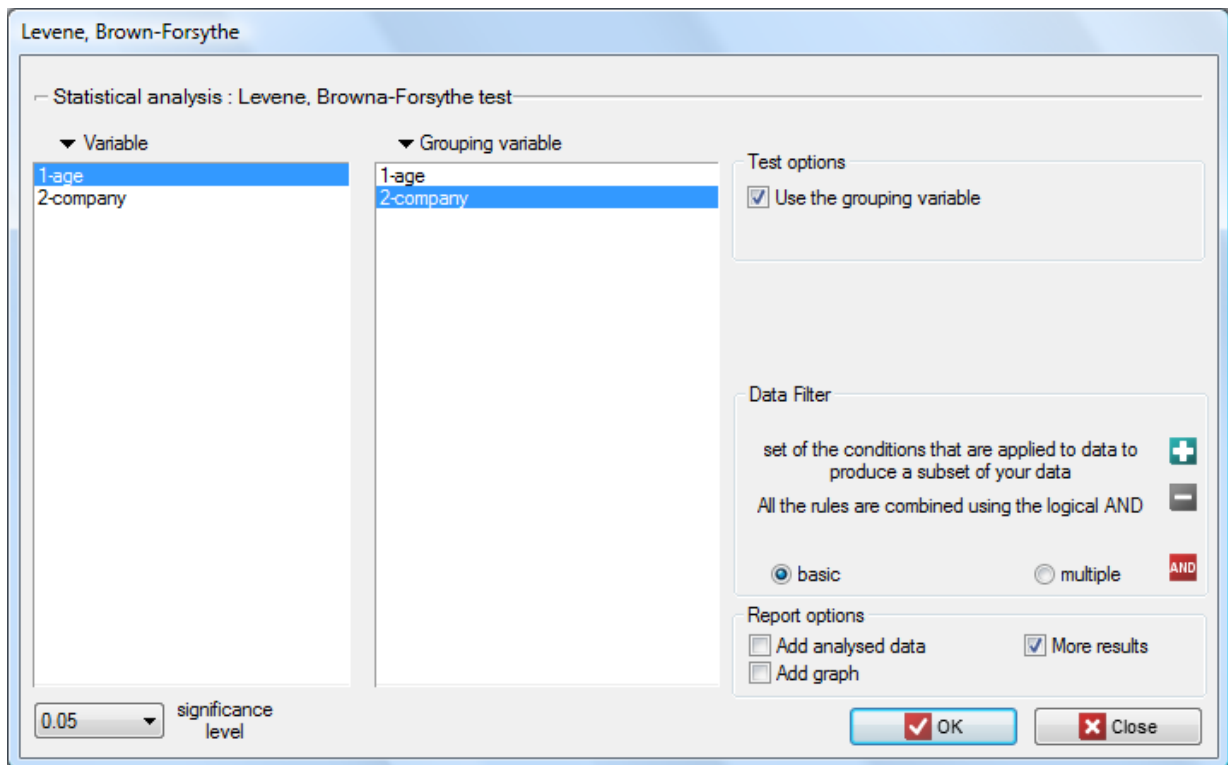
The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha & \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

Note

The Brown-Forsythe test is less sensitive than the Levene test, in terms of an unfulfilled assumption relating to distribution normality.

The settings window with the Levene, Brown-Forsythe tests' can be opened in Statistics menu → Parametric tests → Levene, Brown-Forsythe.



12.1.4 The ANOVA for dependent groups

The single-factor repeated-measures analysis of variance (ANOVA for dependent groups) is used when the measurements of an analysed variable are made several times ($k \geq 2$) each time in different conditions (but we need to assume that the variances of the differences between all the pairs of measurements are pretty close to each other).

This test is used to verify the hypothesis determining the equality of **means** of an analysed variable in several ($k \geq 2$) populations.

Basic assumptions:

- measurement on an **interval scale**,
- the **normal distribution** for all variables which are the differences of measurement pairs (or the normal distribution for an analysed variable in each measurement),
- a **dependent model**.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \mu_1 = \mu_2 = \dots = \mu_k, \\ \mathcal{H}_1 : & \text{not all } \mu_j \text{ are equal } (j = 1, 2, \dots, k),\end{aligned}$$

where:

$\mu_1, \mu_2, \dots, \mu_k$ – means for an analysed features, in the following measurements from the examined population.

The test statistic is defined by:

$$F = \frac{MS_{BC}}{MS_{res}}$$

where:

$$MS_{BC} = \frac{SS_{BC}}{df_{BC}} - \text{mean square between-conditions},$$

$$MS_{res} = \frac{SS_{res}}{df_{res}} - \text{mean square residual},$$

$$SS_{BC} = \sum_{j=1}^k \left(\frac{(\sum_{i=1}^n x_{ij})^2}{n} \right) - \frac{(\sum_{j=1}^k \sum_{i=1}^n x_{ij})^2}{N} - \text{between-conditions sum of squares},$$

$$SS_{res} = SS_T - SS_{BS} - SS_{BC} - \text{residual sum of squares},$$

$$SS_T = \left(\sum_{j=1}^k \sum_{i=1}^n x_{ij}^2 \right) - \frac{(\sum_{j=1}^k \sum_{i=1}^n x_{ij})^2}{N} - \text{total sum of squares},$$

$$SS_{BS} = \sum_{i=1}^n \left(\frac{(\sum_{j=1}^k x_{ij})^2}{k} \right) - \frac{(\sum_{j=1}^k \sum_{i=1}^n x_{ij})^2}{N} - \text{between-subjects sum of squares},$$

$$df_{BC} = k - 1 - \text{between-conditions degrees of freedom},$$

$$df_{res} = df_T - df_{BC} - df_{BS} - \text{residual degrees of freedom},$$

$$df_T = N - 1 - \text{total degrees of freedom},$$

$$df_{BS} = n - 1 - \text{between-subjects degrees of freedom},$$

$$N = nk,$$

$$n - \text{sample size},$$

$$x_{ij} - \text{values of the variable from } i \text{ subjects } (i = 1, 2, \dots, n) \text{ in } j \text{ conditions } (j = 1, 2, \dots, k).$$

The test statistic has the **F Snedecor distribution** with df_{BC} and df_{res} degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

The POST-HOC tests

Introduction to the **contrasts and the POST-HOC tests** was performed in the 12.1.2 unit, which relates to the one-way analysis of variance.

The LSD Fisher test

For simple and complex comparisons (frequency in particular measurements is always the same).

Hypotheses:

Example - **simple comparisons** (comparison of 2 selected means):

$$\begin{aligned} \mathcal{H}_0 : \mu_j &= \mu_{j+1}, \\ \mathcal{H}_1 : \mu_j &\neq \mu_{j+1}. \end{aligned}$$

(i) The value of the critical difference is calculated by using the following formula:

$$CD = \sqrt{F_{\alpha,1,df_{res}}} \cdot \sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n} \right) MS_{res}},$$

where:

$F_{\alpha,1,df_{res}}$ - is the **critical value** (statistic) of the **F Snedecor distribution** for a given significance level α and degrees of freedom, adequately: 1 and df_{res} .

(ii) The test statistic is defined by:

$$t = \frac{\sum_{j=1}^k c_j \bar{x}_j}{\sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n}\right) MS_{res}}}.$$

The test statistic has the **t-Student distribution** with df_{res} degrees of freedom.

The Scheffe test

For simple comparisons (frequency in particular measurements is always the same).

(i) The value of the critical difference is calculated by using the following formula:

$$CD = \sqrt{F_{\alpha,df_{BC},df_{res}}} \cdot \sqrt{(k-1) \left(\sum_{j=1}^k \frac{c_j^2}{n}\right) MS_{res}},$$

where:

$F_{\alpha,df_{BC},df_{res}}$ - is the **critical value** (statistic) of the **F Snedecor distribution** for a given significance level α and df_{BC} and df_{res} degrees of freedom.

(ii) The test statistic is defined by:

$$F = \frac{\left(\sum_{j=1}^k c_j \bar{x}_j\right)^2}{(k-1) \left(\sum_{j=1}^k \frac{c_j^2}{n}\right) MS_{res}}.$$

The test statistic has the **F Snedecor distribution** with df_{BC} and df_{ref} degrees of freedom.

The Tukey test.

For simple comparisons (frequency in particular measurements is always the same).

(i) The value of the critical difference is calculated by using the following formula:

$$CD = \frac{\sqrt{2} \cdot q_{\alpha,df_{WG},k} \cdot \sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n}\right) MS_{res}}}{2},$$

where:

$q_{\alpha,df_{res},k}$ - is the **critical value** (statistic) of the studentized range distribution for a given significance level α and df_{res} and k degrees of freedom.

(ii) The test statistic is defined by:

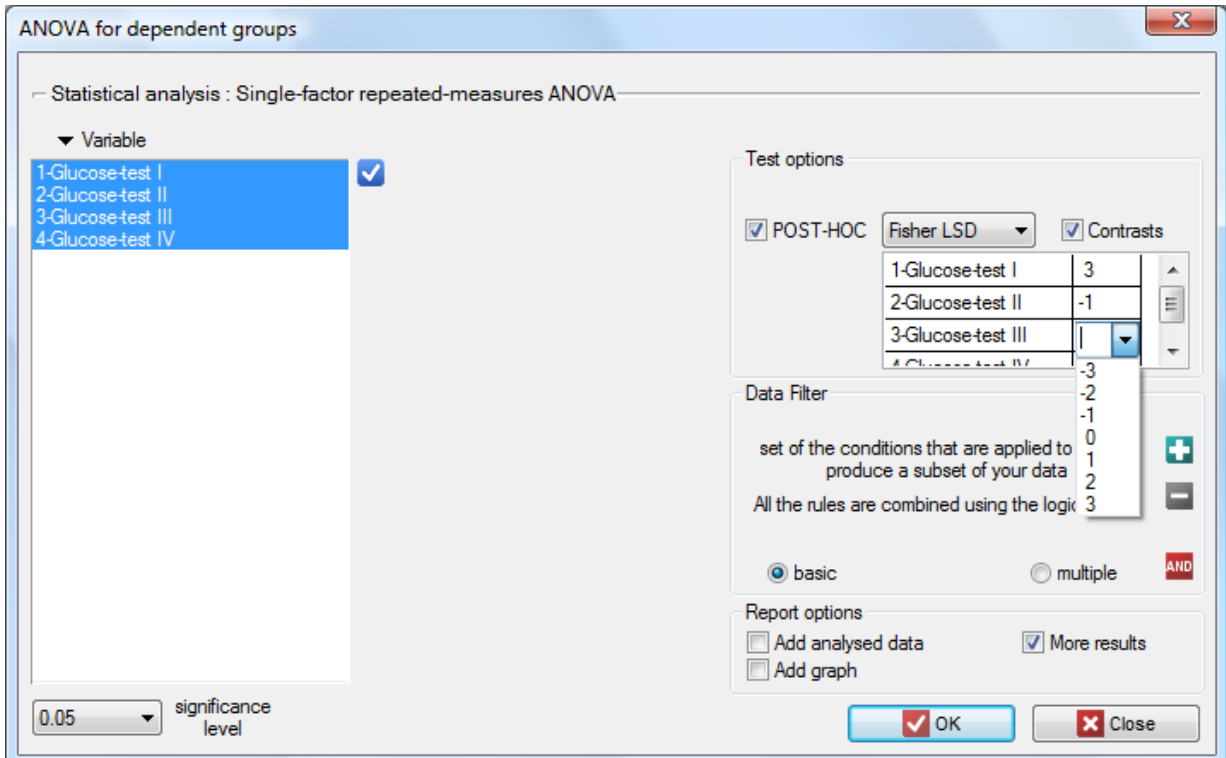
$$q = \sqrt{2} \frac{\sum_{j=1}^k c_j \bar{x}_j}{\sqrt{\left(\sum_{j=1}^k \frac{c_j^2}{n}\right) MS_{res}}}.$$

The test statistic has the studentized range distribution with df_{res} and k degrees of freedom.

Info.

The algorithm for calculating the p value and statistic of the studentized range distribution in PQStat is based on the Lund works (1983)[54]. Other applications or web pages may calculate a little bit different values than PQStat, because they may be based on less precised or more restrictive algorithms (Copenhaver and Holland (1988), Gleason (1999)).

The settings window with the Single-factor repeated-measures ANOVA can be opened in Statistics menu→Parametric tests→ANOVA for dependent groups or in [Wizard](#).



ANOVA for dependent groups

Statistical analysis : Single-factor repeated-measures ANOVA

Variable

- 1-Glucose-test I
- 2-Glucose-test II
- 3-Glucose-test III
- 4-Glucose-test IV

Test options

☒ POST-HOC Fisher LSD ☒ Contrasts

1-Glucose-test I	3
2-Glucose-test II	-1
3-Glucose-test III	1
4-Glucose-test IV	-3

Data Filter

set of the conditions that are applied to produce a subset of your data

All the rules are combined using the logic

☒ basic ☐ multiple AND

Report options

☐ Add analysed data ☒ More results

☐ Add graph

0.05 significance level

OK Close

12.2 NONPARAMETRIC TESTS

12.2.1 The Kruskal-Wallis ANOVA

The Kruskal-Wallis one-way analysis of variance by ranks (Kruskal 1952 [46]; Kruskal and Wallis 1952 [47]) is an extension of the [U-Mann-Whitney test](#) on more than two populations. This test is used to verify the hypothesis determining insignificant differences between medians of the analysed variable in ($k \geq 2$) populations (but you need to assume, that the variable distributions are similar).

Basic assumptions:

- measurement on an [ordinal scale](#) or on an [interval scale](#),
- an [independent model](#).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \theta_1 = \theta_2 = \dots = \theta_k, \\ \mathcal{H}_1 : & \text{not all } \theta_j \text{ are equal } (j = 1, 2, \dots, k),\end{aligned}$$

where:

$\theta_1, \theta_2, \dots, \theta_k$ medians of the analysed variable of each population.

The test statistic is defined by:

$$H = \frac{1}{C} \left(\frac{12}{N(N+1)} \sum_{j=1}^k \left(\frac{(\sum_{i=1}^{n_j} R_{ij})^2}{n_j} \right) - 3(N+1) \right),$$

where:

$$N = \sum_{j=1}^k n_j,$$

n_j – samples sizes ($j = 1, 2, \dots, k$),

R_{ij} – ranks ascribed to the values of a variable for ($i = 1, 2, \dots, n_j$), ($j = 1, 2, \dots, k$),

$$C = 1 - \frac{\sum (t^3 - t)}{N^3 - N} - \text{correction for ties},$$

t – number of cases included in a tie.

The formula for the test statistic H includes the correction for ties C . This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $C = 1$).

The H statistic asymptotically (for large sample sizes) has the χ^2 [distribution](#) with the number of degrees of freedom calculated using the formula: $df = (k - 1)$.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned}\text{if } p &\leq \alpha \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p &> \alpha \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The POST-HOC tests

Introduction to [the contrasts and the POST-HOC tests](#) was performed in the 12.1.2 unit, which relates to the one-way analysis of variance.

The Dunn test

For simple comparisons, equal-size groups as well as unequal-size groups.

Hypotheses:

Example - **simple comparisons** (comparison of 2 selected medians):

$$\begin{aligned}\mathcal{H}_0 &: \theta_j = \theta_{j+1}, \\ \mathcal{H}_1 &: \theta_j \neq \theta_{j+1}.\end{aligned}$$

(i) The value of critical difference is calculated by using the following formula:

$$CD = Z_{\frac{\alpha}{c}} \sqrt{\frac{N(N+1)}{12} \left(\sum_{j=1}^k \frac{c_j^2}{n_j} \right)},$$

where:

$Z_{\frac{\alpha}{c}}$ - is the **critical value** (statistic) of the normal distribution for a given significance level α corrected on the number of possible **simple comparisons** c .

(ii) The test statistic is defined by:

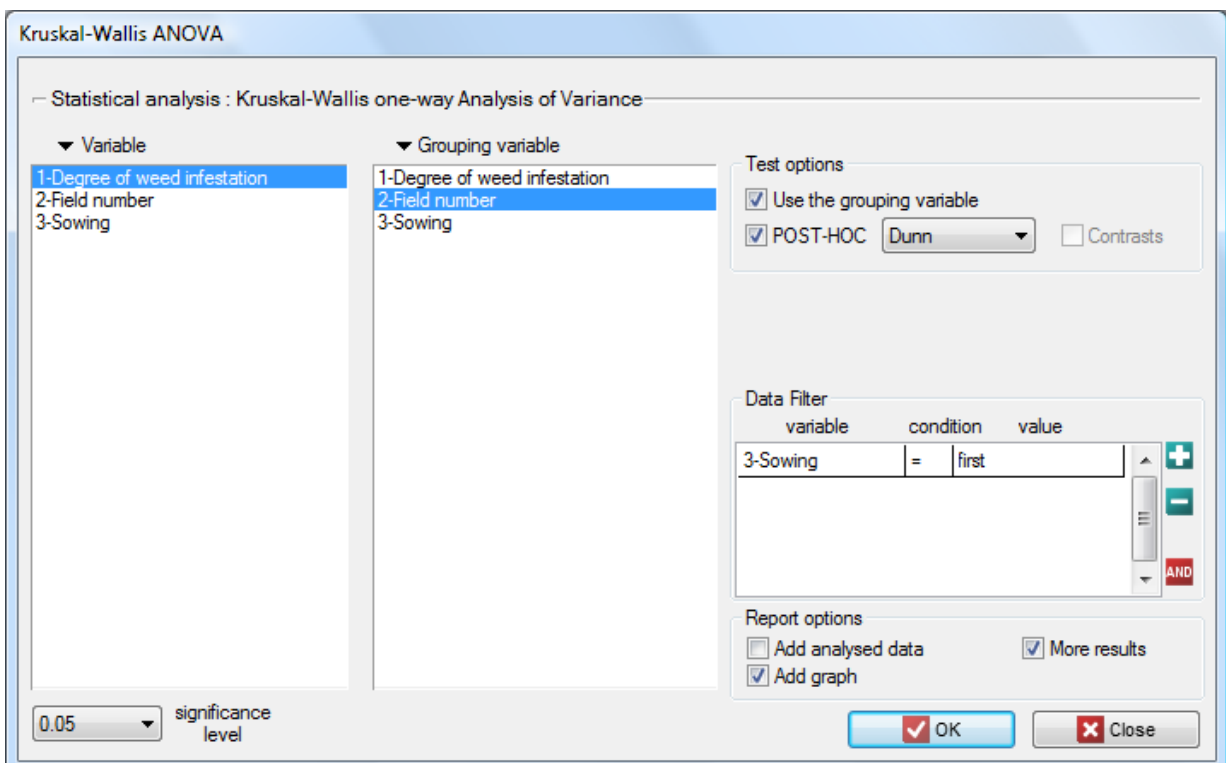
$$Z = \frac{\sum_{j=1}^k c_j \bar{R}_j}{\sqrt{\frac{N(N+1)}{12} \left(\sum_{j=1}^k \frac{c_j^2}{n_j} \right)}},$$

where:

\bar{R}_j – mean of the ranks of the j -th group, for $(j = 1, 2, \dots, k)$,

The test statistic asymptotically (for large sample sizes) has the **normal distribution**, and the **p value** is corrected on the number of possible **simple comparisons** c .

The settings window with the Kruskal-Wallis ANOVA can be opened in Statistics menu→NonParametric tests (ordered categories)→Kruskal-Wallis ANOVA or in **Wizard**.



Kruskal-Wallis ANOVA

Statistical analysis : Kruskal-Wallis one-way Analysis of Variance

Variable

- 1-Degree of weed infestation
- 2-Field number
- 3-Sowing

Grouping variable

- 1-Degree of weed infestation
- 2-Field number
- 3-Sowing

Test options

- ☒ Use the grouping variable
- ☒ POST-HOC **Dunn**
- ☐ Contrasts

Data Filter

variable	condition	value
3-Sowing	=	first

Report options

- ☐ Add analysed data
- ☒ Add graph
- ☒ More results

0.05 significance level

OK Close

12.2.2 The Friedman ANOVA

The Friedman repeated measures analysis of variance by ranks – the Friedman ANOVA - was described by Friedman (1937)[33]. This test is used when the measurements of an analysed variable are made several times ($k \geq 2$) each time in different conditions. It is also used when we have rankings coming from different sources (from different judges) and concerning a few ($k \geq 2$) objects, but we want to assess the grade of the rankings agreement.

Basic assumptions:

- measurement on an **ordinal scale** or on an **interval scale**,
- a **dependent model**.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \theta_1 = \theta_2 = \dots = \theta_k, \\ \mathcal{H}_1 : & \text{not all } \theta_j \text{ are equal } (j = 1, 2, \dots, k),\end{aligned}$$

where:

$\theta_1, \theta_2, \dots, \theta_k$ medians for an analysed features, in the following measurements from the examined population.

The test statistic is defined by:

$$\chi_r^2 = \frac{1}{C} \left(\frac{12}{nk(k+1)} \left(\sum_{j=1}^k \left(\sum_{i=1}^n R_{ij} \right)^2 \right) - 3n(k+1) \right),$$

where:

n – sample size,

R_{ij} – ranks ascribed to the following measurements ($j = 1, 2, \dots, k$), separately for the analysed objects ($i = 1, 2, \dots, n$),

$C = 1 - \frac{\sum(t^3 - t)}{n(k^3 - k)}$ – correction for **ties**,

t – number of cases included in a tie.

The formula for the test statistic χ_r^2 includes the correction for ties C . This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $C = 1$).

The χ_r^2 statistic asymptotically (for large sample size) has the **χ^2 distribution** with the number of degrees of freedom calculated using the formula: $df = (k - 1)$.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha & \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The POST-HOC tests

Introduction to **the contrasts and the POST-HOC tests** was performed in the 12.1.2 unit, which relates to the one-way analysis of variance.

The Dunn test

For simple comparisons (frequency in particular measurements is always the same).

Hypotheses:

Example - **simple comparisons** (comparison of 2 selected medians):

$$\begin{aligned}\mathcal{H}_0 &: \theta_j = \theta_{j+1}, \\ \mathcal{H}_1 &: \theta_j \neq \theta_{j+1}.\end{aligned}$$

(i) The value of critical difference is calculated by using the following formula:

$$NIR = Z_{\frac{\alpha}{c}} \sqrt{\frac{k(k+1)}{6n}},$$

where:

$Z_{\frac{\alpha}{c}}$ - is the **critical value** (statistic) of the normal distribution for a given significance level α corrected on the number of possible **simple comparisons** c .

(ii) The test statistic is defined by:

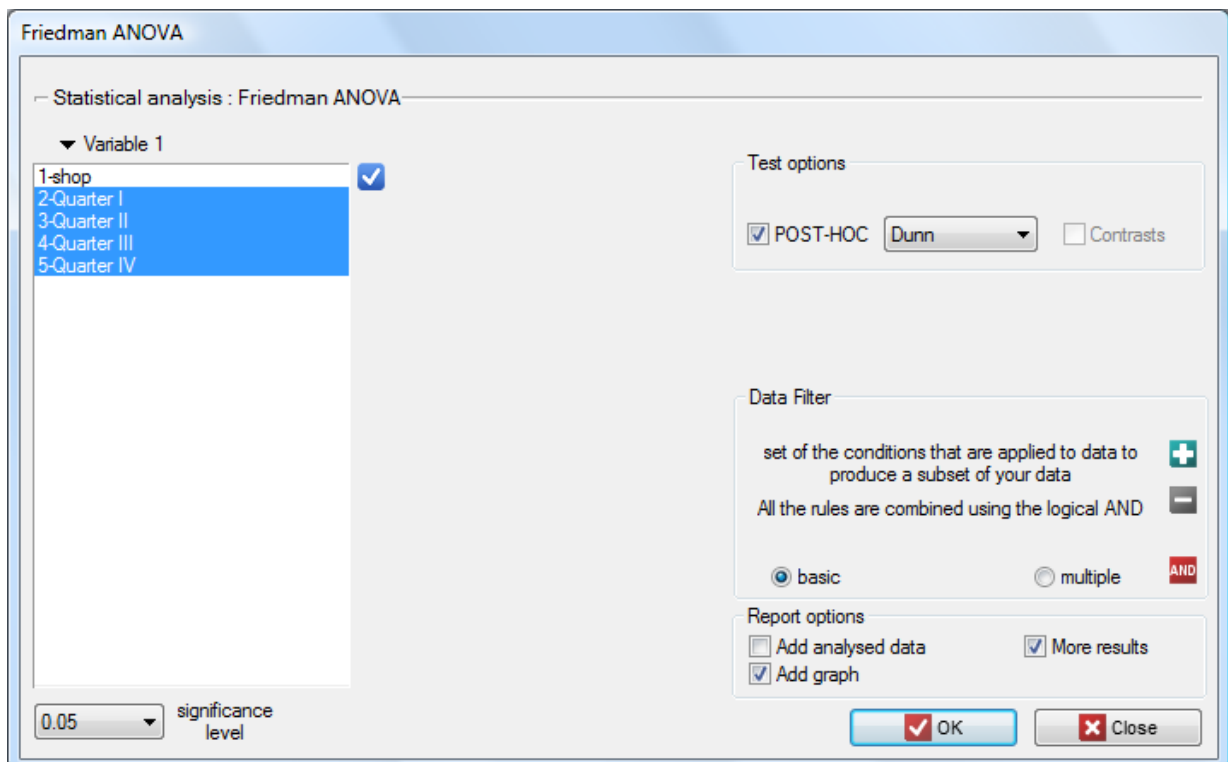
$$Z = \frac{\sum_{j=1}^k c_j R_j}{\sqrt{\frac{k(k+1)}{6n}}},$$

where:

\bar{R}_j – mean of the ranks of the j -th measurement, for $(j = 1, 2, \dots, k)$,

The test statistic asymptotically (for large sample size) has **normal distribution**, and the **p value** is corrected on the number of possible **simple comparisons** c .

The settings window with the Friedman ANOVA can be opened in Statistics menu → NonParametric tests (ordered categories) → Friedman ANOVA or in **Wizard**.



EXAMPLE 12.2. (chocolate bar.pqs file)

Quarterly sale of some chocolate bar was measured in 14 randomly chosen supermarkets. The study was started in January and finished in December. During the second quarter, the billboard campaign was in full swing. Let's check if the campaign had an influence on the advertised chocolate bar sale.

Shop	Quarter I	Quarter II	Quarter III	Quarter IV
SK1	3415	4556	5772	5432
SK2	1593	1937	2242	2794
SK3	1976	2056	2240	2085
SK4	1526	1594	1644	1705
SK5	1538	1634	1866	1769
SK6	983	1086	1135	1177
SK7	1050	1209	1245	977
SK8	1861	2087	2054	2018
SK9	1714	2415	2361	2424
SK10	1320	1621	1624	1551
SK11	1276	1377	1522	1412
SK12	1263	1279	1350	1490
SK13	1271	1417	1583	1513
SK14	1436	1310	1357	1468

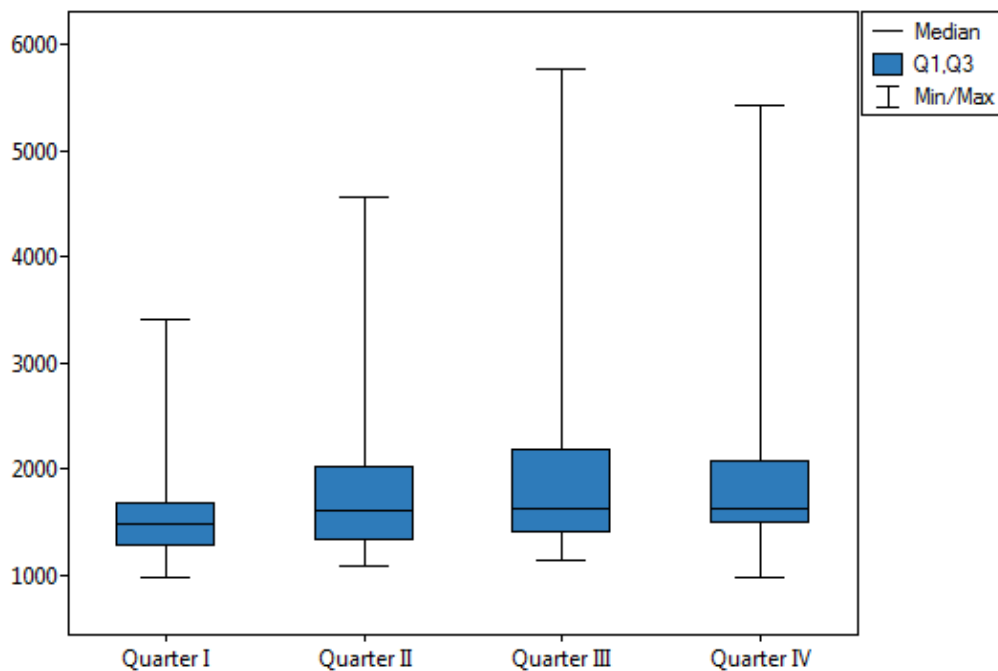
Hypotheses:

- \mathcal{H}_0 : there is a lack of significant difference in sale values, in the compared quarters, in the population represented by the whole sample,
 \mathcal{H}_1 : the difference in sale values, between at least 2 quarters, is significant, in the population represented by the whole sample.

Friedman ANOVA	
Analysis time	0.03sec.
Analysed variables	Quarter I,Quarter II,Quarter III,Quarter I
Significance level	0.05
Group name	Quarter I
Group size	14
Sum of the ranks for group	17
Mean of the ranks for the group	1.214286
Group median	1481
Group name	Quarter II
Group size	14
Sum of the ranks for group	32
Mean of the ranks for the group	2.285714
Group median	1607.5
Group name	Quarter III
Group size	14
Sum of the ranks for group	47
Mean of the ranks for the group	3.357143
Group median	1634
Group name	Quarter IV
Group size	14
Sum of the ranks for group	44
Mean of the ranks for the group	3.142857
Group median	1628
Degrees of freedom	3
Chi2 statistic (adjusted for ties)	23.914286
p-value	0.000026

Comparing the $p = 0,000026$ with the significance level $\alpha = 0.05$, we state that the chocolate bar sale is not the same in each quarter. The POST-HOC analysis indicates the difference in the sale in quarters I/III and I/IV.

POST-HOC (Dunn)				
	Quarter I	Quarter II	Quarter II	Quarter IV
Difference of the means				
Quarter I		1.07143	2.14286	1.92857
Quarter II	1.07143		1.07143	0.85714
Quarter III	2.14286	1.07143		0.21429
Quarter IV	1.92857	0.85714	0.21429	
CD				
Quarter I		1.28734	1.28734	1.28734
Quarter II	1.28734		1.28734	1.28734
Quarter III	1.28734	1.28734		1.28734
Quarter IV	1.28734	1.28734	1.28734	
Statistic Z				
Quarter I		2.19578	4.39155	3.9524
Quarter II	2.19578		2.19578	1.75662
Quarter III	4.39155	2.19578		0.43916
Quarter IV	3.9524	1.75662	0.43916	
p-value				
Quarter I		0.16865	0.00007	0.00046
Quarter II	0.16865		0.16865	0.4739
Quarter III	0.00007	0.16865		1
Quarter IV	0.00046	0.4739	1	



12.2.3 The Chi-square test for multidimensional contingency tables

The χ^2 test for multidimensional contingency tables is an extension to the χ^2 test for $(R \times C)$ tables for more than two features.

Basic assumptions:

- measurement on a **nominal scale** (alternatively: an **ordinal scale** or an **interval scale**),

- an **independent model**,
- large **expected frequencies** (according to the Cochran interpretation (1952)[20], none of these expected frequencies can be < 1 and no more than 20% of the expected frequencies can be < 5).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 &: O_{ij...} = E_{ij...} \text{ for all categories,} \\ \mathcal{H}_1 &: O_{ij...} \neq E_{ij...} \text{ for at least one category,}\end{aligned}$$

where:

$O_{ij...}$ and $E_{ij...}$ – **observed frequencies** in a contingency table and the corresponding **expected frequencies**.

The test statistic is defined by:

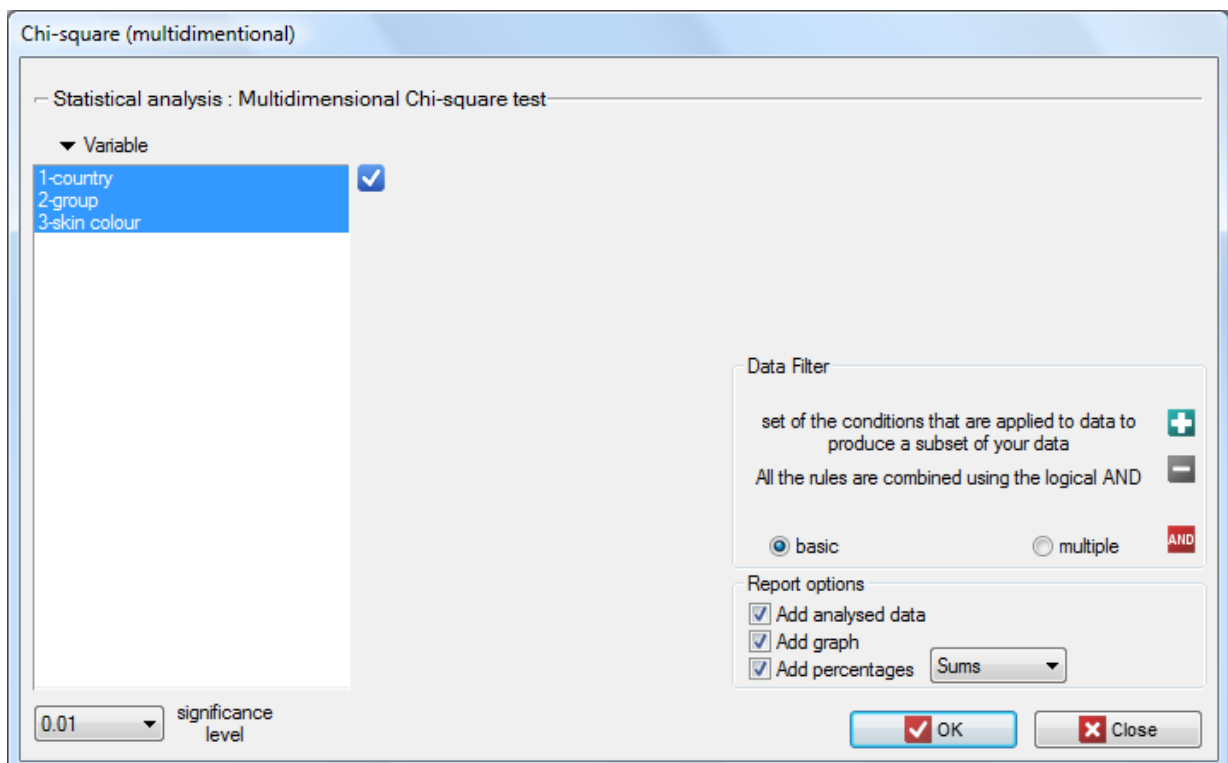
$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \sum \dots \sum \frac{(O_{ij...} - E_{ij...})^2}{E_{ij...}}.$$

This statistic asymptotically (for large expected frequencies) has the **χ^2 distribution** with a number of degrees of freedom calculated using the formula: $df = (r - 1)(c - 1)(l - 1) + (r - 1)(c - 1) + (r - 1)(l - 1) + (c - 1)(l - 1)$ - for 3-dimensional tables.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level **α** :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The settings window with the Chi-square (multidimensional) test can be opened in Statistics menu → NonParametric tests (unordered categories) → Chi-square (multidimensional) or in **Wizard**.



Note

This test can be calculated only on the basis of [raw data](#).

12.2.4 The Q-Cochran ANOVA

The Q-Cochran analysis of variance, based on the Q-Cochran test, is described by Cochran (1950)[19]. This test is an extended [McNemar](#) test for $k \geq 2$ dependent groups. It is used in hypothesis verification about symmetry between several measurements $X^{(1)}, X^{(2)}, \dots, X^{(k)}$ for the X feature. The analysed feature can have only 2 values - for the analysis, there are ascribed to them the numbers: 1 and 0.

Basic assumptions:

- measurement on a [nominal scale](#) (dichotomous variables – it means the variables of two categories),
- a [dependent model](#).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 &: \text{all the "incompatible" observed frequencies are equal,} \\ \mathcal{H}_1 &: \text{not all the "incompatible" observed frequencies are equal,}\end{aligned}$$

where:

"incompatible" observed frequencies – the [observed frequencies](#) calculated when the value of the analysed feature is different in several measurements.

The test statistic is defined by:

$$Q = \frac{(k-1)(kC - T^2)}{kT - R}$$

where:

$$T = \sum_{i=1}^n \sum_{j=1}^k x_{ij},$$

$$R = \sum_{i=1}^n \left(\sum_{j=1}^k x_{ij} \right)^2,$$

$$C = \sum_{j=1}^k \left(\sum_{i=1}^n x_{ij} \right)^2,$$

x_{ij} – the value of j -th measurement for i -th object (so 0 or 1).

This statistic asymptotically (for large sample size) has the χ^2 [distribution](#) with a number of degrees of freedom calculated using the formula: $df = k - 1$.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The POST-HOC tests

Introduction to [the contrasts and the POST-HOC tests](#) was performed in the 12.1.2 unit, which relates to the one-way analysis of variance.

The Dunn test

For simple comparisons (frequency in particular measurements is always the same).

Hypotheses:

Example - **simple comparisons** (for the difference in proportion in a one chosen pair of measurements):

\mathcal{H}_0 : the chosen "incompatible" observed frequencies are equal,
 \mathcal{H}_1 : the chosen "incompatible" observed frequencies are different.

(i) The value of critical difference is calculated by using the following formula:

$$NIR = Z_{\frac{\alpha}{c}} \sqrt{2 \frac{kT - R}{n^2 k(k-1)}},$$

where:

$Z_{\frac{\alpha}{c}}$ - is the **critical value** (statistic) of the normal distribution for a given significance level α corrected on the number of possible **simple comparisons** c .

(ii) The test statistic is defined by:

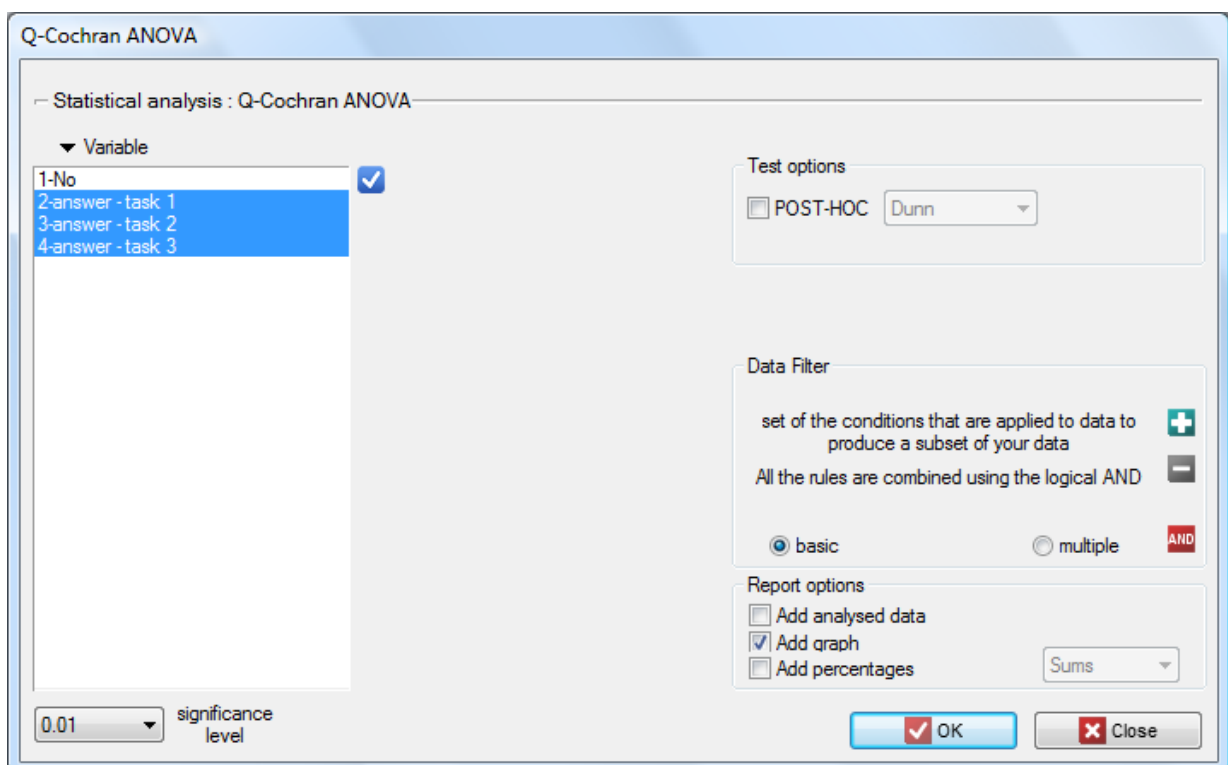
$$Z = \frac{\sum_{j=1}^k c_j p_j}{\sqrt{2 \frac{kT - R}{n^2 k(k-1)}}},$$

where:

p_j - the proportion j -th measurement ($j = 1, 2, \dots, k$),

The test statistic asymptotically (for large sample size) has the **normal distribution**, and the **p value** is corrected on the number of possible **simple comparisons** c .

The settings window with the Cochran Q ANOVA can be opened in Statistics menu → NonParametric tests (unordered categories) → Cochran Q ANOVA or in **Wizard**.



Note

This test can be calculated only on the basis of **raw data**.

EXAMPLE 12.3. (test.pqs file)

We want to compare the difficulty of 3 test questions. To do this, we select a sample of 20 people from the analysed population. Every person from the sample answers 3 test questions. Next, we check the correctness of answers (an answer can be correct or wrong). In the table, there are following scores:


No.	question 1 answer	question 2 answer	question 3 answer
1	correct	correct	wrong
2	wrong	correct	wrong
3	correct	correct	correct
4	wrong	correct	wrong
5	wrong	correct	wrong
6	wrong	correct	correct
7	wrong	wrong	wrong
8	wrong	correct	wrong
9	correct	correct	wrong
10	wrong	correct	wrong
11	wrong	wrong	wrong
12	wrong	wrong	correct
13	wrong	correct	wrong
14	wrong	wrong	correct
15	correct	wrong	wrong
16	wrong	wrong	wrong
17	wrong	correct	wrong
18	wrong	correct	wrong
19	wrong	wrong	wrong
20	correct	correct	wrong

Hypotheses:

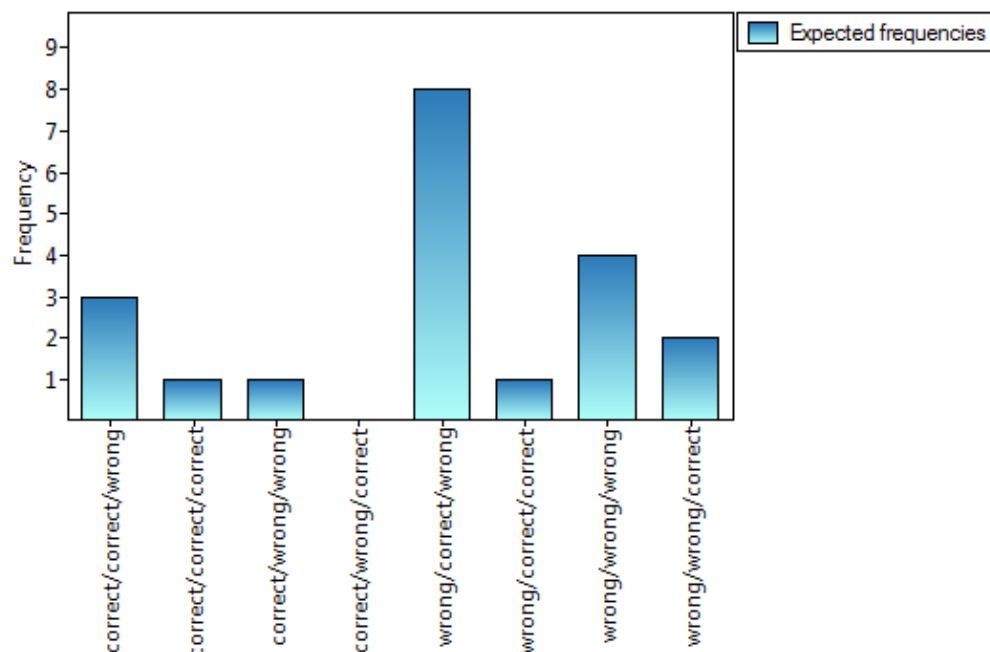
\mathcal{H}_0 : The individual questions received the same number of correct answers, in the analysed population,

\mathcal{H}_1 : There are different numbers of correct and wrong answers in individual test questions, in the analysed population.

Cochran Q ANOVA	
Analysis time	0.03sec.
Analysed variables	answer - task 1,answer - task 2,answer - ta
Significance level	0.05
Size	20
Degrees of freedom	2
Statistic Q	9.733333
p-value	0.007699

Comparing the p value $p = 0.007699$ with the significance level $\alpha = 0.05$ we conclude that individual test questions have different difficulty levels. We resume the analysis to perform POST-HOC test by clicking , and in the test option window, we select POST-HOC Dunn.

POST-HOC (Dunn)			
	answer - t	answer - t	answer - t
Difference of the mean:			
answer - task 1		0.4	0.05
answer - task 2	0.4		0.45
answer - task 3	0.05	0.45	
CD			
answer - task 1		0.4641	0.4641
answer - task 2	0.4641		0.4641
answer - task 3	0.4641	0.4641	
Statistic Z			
answer - task 1		2.52982	0.31623
answer - task 2	2.52982		2.84605
answer - task 3	0.31623	2.84605	
p-value			
answer - task 1		0.03424	1
answer - task 2	0.03424		0.01328
answer - task 3	1	0.01328	



The carried out POST-HOC analysis indicates that there are differences between the 2-nd and 1-st question and between questions 2-nd and 3-th. The difference is because the second question is easier than the first and the third ones (the number of correct answers the first question is higher).

13 STRATIFIED ANALYSIS

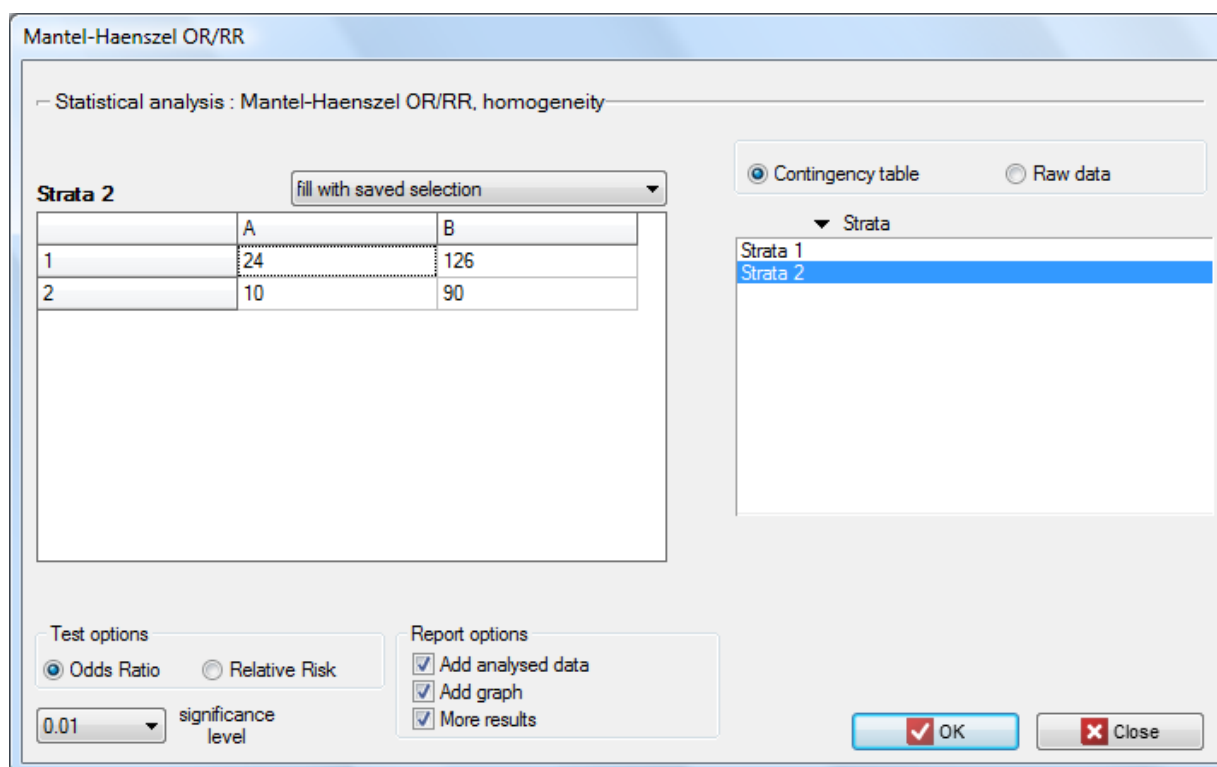
13.1 THE MANTEL - HAENSZEL METHOD FOR SEVERAL 2x2 TABLES

The Mantel-Haenszel method for 2×2 tables proposed by Mantel and Haenszel (1959)[56] then it was extended by Mantel (1963)[57]. A wider review the development of these methods was carried out i.a. by Newman (2001)[66].

This method can be used in analysis 2×2 tables, that occur in several ($w \geq 2$) stratas constructed by confounding variable. For the next stratas ($s = 1, \dots, w$) the 2×2 contingency tables for observed frequencies are created:

Observed frequencies s -th strata ($O_{ij}^{(s)}$)		Analysed phenomenon (illness)		
		occurs (case)	not occurs (control)	Total
Risk factor	exposed	$O_{11}^{(s)}$	$O_{12}^{(s)}$	$O_{11}^{(s)} + O_{12}^{(s)}$
	unexposed	$O_{21}^{(s)}$	$O_{22}^{(s)}$	$O_{21}^{(s)} + O_{22}^{(s)}$
	Total	$O_{11}^{(s)} + O_{21}^{(s)}$	$O_{12}^{(s)} + O_{22}^{(s)}$	$n^{(s)} = O_{11}^{(s)} + O_{12}^{(s)} + O_{21}^{(s)} + O_{22}^{(s)}$

The settings window with the Mantel–Haenszel OR/RR can be opened in Statistics menu → Stratified analysis → Mantel–Haenszel OR/RR.



13.1.1 The Mantel-Haenszel odds ratio

If all tables (created by individual stratas) are homogeneous (the χ^2 test of homogeneity for the OR can check this condition), then, on the basis of these tables, the pooled odds ratio with the confidence interval can be designated. Such odds ratio, is a weighted mean for an odds ratio designated for the individual stratas. The usage of the weighted method, proposed by Mantel and Haenszel allows to include the contribution of the strata weights. Each strata has an influence on the pooled odds ratio (the greater size of the strata, the greater weight and the greater influence on the pooled odds ratio).

Weights for individual stratas are designated according to the following formula:

$$g^{(s)} = \frac{O_{21}^{(s)} \cdot O_{12}^{(s)}}{n^{(s)}},$$

and the **Mantel-Haenszel odds ratio**:

$$OR_{MH} = \frac{R}{S},$$

where:

$$R = \sum_{s=1}^w \frac{O_{11}^{(s)} \cdot O_{22}^{(s)}}{n^{(s)}},$$

$$S = \sum_{s=1}^w g^{(s)}.$$

The confidence interval for $\log OR_{MH}$ is designated on the basis of the standard error (RGB – Robins-Breslow-Greenland[70][71]) calculated according to the following formula:

$$SE_{MH} = \sqrt{\frac{T}{2R^2} + \frac{U+Y}{2RS} + \frac{W}{2S^2}},$$

where:

$$T = \sum_{s=1}^w T^{(s)}, \quad T^{(s)} = \frac{O_{11}^{(s)} \cdot O_{22}^{(s)} \cdot (O_{11}^{(s)} + O_{22}^{(s)})}{(n^{(s)})^2},$$

$$U = \sum_{s=1}^w U^{(s)}, \quad U^{(s)} = \frac{O_{21}^{(s)} \cdot O_{12}^{(s)} \cdot (O_{11}^{(s)} + O_{22}^{(s)})}{(n^{(s)})^2},$$

$$Y = \sum_{s=1}^w Y^{(s)}, \quad Y^{(s)} = \frac{O_{11}^{(s)} \cdot O_{22}^{(s)} \cdot (O_{21}^{(s)} + O_{12}^{(s)})}{(n^{(s)})^2},$$

$$W = \sum_{s=1}^w W^{(s)}, \quad W^{(s)} = \frac{O_{21}^{(s)} \cdot O_{12}^{(s)} \cdot (O_{21}^{(s)} + O_{12}^{(s)})}{(n^{(s)})^2}.$$

The Mantel-Haenszel χ^2 test for the OR_{MH}

The Mantel-Haenszel Chi-square test for the OR_{MH} is used in the hypothesis verification about the significance of designated odds ratio (OR_{MH}). It should be calculated for large frequencies, i.e. when both conditions of the so-called "rule 5" are satisfied:

- $\min(O_{11}^{(s)} + O_{12}^{(s)}, O_{11}^{(s)} + O_{21}^{(s)}) - \sum_{s=1}^w E_{11}^{(s)} \geq 5$ for all the stratas $s = 1, 2, \dots, w$,
- $\max(0, O_{11}^{(s)} - O_{22}^{(s)}) \geq 5$ for all the stratas $s = 1, 2, \dots, w$.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : & \quad OR_{MH} = 1, \\ \mathcal{H}_1 : & \quad OR_{MH} \neq 1. \end{aligned}$$

The test statistic is defined by:

$$\chi_{MH}^2 = \frac{\left(\sum_{s=1}^w O_{11}^{(s)} - \sum_{s=1}^w E_{11}^{(s)} \right)^2}{V},$$

where:

$E_{11}^{(s)} = \frac{(O_{11}^{(s)} + O_{21}^{(s)}) (O_{11}^{(s)} + O_{12}^{(s)})}{n^{(s)}}$ are the expected frequencies in the first contingency table cell, for the individual stratas $s = 1, 2, \dots, w$,

$$V = \sum_{s=1}^w V^{(s)},$$

$$V^{(s)} = \frac{(O_{11}^{(s)} + O_{12}^{(s)}) (O_{21}^{(s)} + O_{22}^{(s)}) (O_{11}^{(s)} + O_{21}^{(s)}) (O_{12}^{(s)} + O_{22}^{(s)})}{(n^{(s)})^2 (n^{(s)} - 1)}.$$

This statistic asymptotically (for large frequencies) has the χ^2 distribution with 1 degree of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

The χ^2 test of homogeneity for the OR

The Chi-square test of homogeneity for the OR is used in the hypothesis verification that the variable, creating stratas, is the modifying effect, i.e. it influences on the designated odds ratio in the manner that, the odds ratios are significant different for individual stratas.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : & \text{OR}_{MH} = \text{OR}^{(s)}, \text{ for all the stratas } s = 1, 2, \dots, w, \\ \mathcal{H}_1 : & \text{OR}_{MH} \neq \text{OR}^{(s)}, \text{ for at least one strata.} \end{aligned}$$

The test statistic (Breslow-Day (1980)[12], Tarone (1985)[13][77]) is defined by:

$$\chi^2 = \sum_{s=1}^w \frac{(O_{11}^{(s)} - E^{(s)})^2}{Var^{(s)}} - \frac{(\sum_{s=1}^w O_{11}^{(s)} - \sum_{s=1}^w E^{(s)})^2}{\sum_{s=1}^w Var^{(s)}}$$

where:

$E^{(s)}$ is solution to the quadratic equation:

$$\frac{E^{(s)} (O_{22}^{(s)} - O_{11}^{(s)} + E^{(s)})}{(O_{11}^{(s)} + O_{21}^{(s)} - E^{(s)}) (O_{11}^{(s)} + O_{12}^{(s)} - E^{(s)})} = \text{OR}_{MH},$$

$$Var^{(s)} = \left(\frac{1}{E^{(s)}} + \frac{1}{O_{22}^{(s)} - O_{11}^{(s)} + E^{(s)}} + \frac{1}{O_{11}^{(s)} + O_{21}^{(s)} - E^{(s)}} + \frac{1}{O_{11}^{(s)} + O_{12}^{(s)} - E^{(s)}} \right)^{-1}.$$

This statistic asymptotically (for large frequencies) has the χ^2 distribution with the number of degrees of freedom calculated using the formula: $df = w - 1$.

The p value, designated on the basis of the test statistic, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

EXAMPLE 13.1. (leptospirosis.pqs file)

The following table presents hypothetical poll results, conducted among inhabitants of a city and village (the village is treated as a risk factor) in West India. The poll aim was to detect risk factors of leptospirosis[9]. The occurrence of leptospirosis antibodies is an indirect evidence about infection.

Observed frequencies O_{ij}		leptospirosis antibodies	
		occur	not occur
place of residence	rural	60	140
	urban	60	140

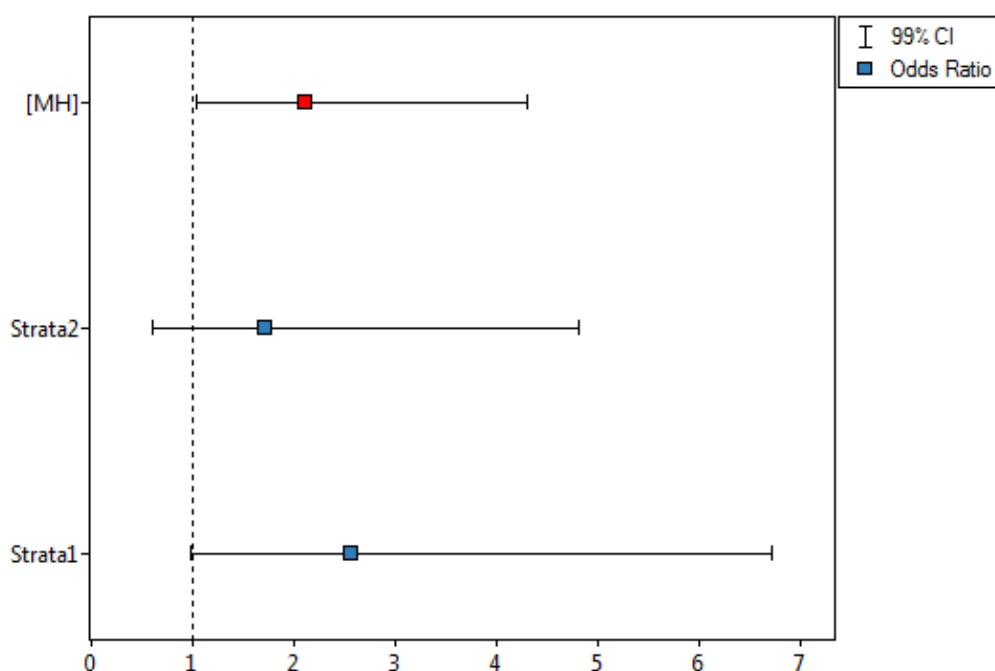
The odds of the occurrence of leptospirosis antibodies, among inhabitants of the city and the village, is the same ($OR=1$). Let's include gender in the analysis and check what odds will be then. The sample has to be divided into 2 stratas, because of gender (they are marked in a file as a [saved selection](#)):

Observed frequencies for men		leptospirosis antibodies	
		occur	not occur
place of residence	rural	36	14
	urban	50	50

Observed frequencies for women		leptospirosis antibodies	
		occur	not occur
place of residence	rural	24	126
	urban	10	90

Gender is associated with both factors (the occurrence of leptospirosis antibodies and the residence in West India). This is a significant factor. Its ignorance can lead to errors in results.

Mantel-Haenszel OR/RR, homogeneity	
Analysis time	0.01sec.
Analysed variables	Contingency table
Significance level	0.05
Size	400
Strata 1	
Odds Ratio	2.571429
-95% CI for the Odds Ratio	1.237622
+95% CI for the Odds Ratio	5.342701
Statistic for the Odds Ratio	2.531365
p-value	0.011362
Strata 2	
Odds Ratio	1.714286
-95% CI for the Odds Ratio	0.781346
+95% CI for the Odds Ratio	3.76117
Statistic for the Odds Ratio	1.344493
p-value	0.178789
Odds Ratio [MH]	2.126374
-95% CI for the Odds Ratio [MH]	1.244338
+95% CI for the Odds Ratio [MH]	3.63363
Degrees of freedom	1
Statistic for the Odds Ratio [MH]	7.81939
p-value	0.005169
Homogeneity of the Odds Ratio	
Degrees of freedom	1
Statistic	0.548611
p-value	0.458886



The odds of the occurrence of leptospirosis antibodies is larger among village inhabitants, both among women (OR[95%CI]=2.57[1.24, 5.34]) and men (OR[95%CI]=1.71[0.78, 3.76]). The tables are homogeneous ($p=0.465049$). Thus, we can use the calculated odds ratio, which is mutual for both tables (OR_{MH} [95%CI]=2.13[1.24, 3.65]). Finally, the obtained result indicates that the odds of the occurrence of leptospirosis antibodies is significantly greater among village inhabitants ($p=0.005169$).

13.1.2 The Mantel-Haenszel relative risk

If all tables (created by individual stratas) are homogeneous (the χ^2 test of homogeneity for the RR), can check this condition), then, on the basis of these tables, the pooled relative risk with the confidence interval can be designated. Such relative risk is a weighted mean for a relative risk designated for the individual stratas. The usage of the weighted method, proposed by Mantel and Haenszel allows to include the contribution of the strata weights. Each strata of the input has an influence on the pooled relative risk construction (the greater size of the strata, the greater weight and the greater influence on the pooled relative risk).

Weights for individual stratas are designated according to the following formula:

$$g^{(s)} = \frac{O_{21}^{(s)} (O_{11}^{(s)} + O_{12}^{(s)})}{n^{(s)}},$$

and the Mantel-Haenszel relative risk:

$$RR_{MH} = \frac{R}{S},$$

where:

$$R = \sum_{s=1}^w \frac{O_{11}^{(s)} (O_{21}^{(s)} + O_{22}^{(s)})}{n^{(s)}},$$

$$S = \sum_{s=1}^w g^{(s)}.$$

The confidence interval for $\log RR_{MH}$ is designated on the basis of the standard error calculated according to the following formula:

$$SE_{MH} = \sqrt{\frac{V}{RS}},$$

where:

$$V = \sum_{s=1}^w V^{(s)},$$

$$V^{(s)} = \frac{(O_{11}^{(s)} + O_{12}^{(s)}) (O_{21}^{(s)} + O_{22}^{(s)}) (O_{11}^{(s)} + O_{21}^{(s)}) - (O_{11}^{(s)} * O_{21}^{(s)} * n^{(s)})}{(n^{(s)})^2}.$$

The Mantel-Haenszel χ^2 test for the RR_{MH}

The Mantel-Haenszel Chi-square test for the RR_{MH} is used in the hypothesis verification about the significance of designated relative risk (RR_{MH}). It should be calculated for large frequencies, in a contingency table.

Hypotheses:

$$\mathcal{H}_0 : RR_{MH} = 1,$$

$$\mathcal{H}_1 : RR_{MH} \neq 1.$$

The test statistic is defined by:

$$\chi_{MH}^2 = \frac{\left(\sum_{s=1}^w O_{11}^{(s)} - \sum_{s=1}^w E_{11}^{(s)} \right)^2}{V},$$

where:

$E_{11}^{(s)} = \frac{(O_{11}^{(s)} + O_{21}^{(s)})(O_{11}^{(s)} + O_{12}^{(s)})}{n^{(s)}}$ are the expected frequencies in the first contingency table cell, for individual stratas $s = 1, 2, \dots, w$.

This statistic asymptotically (for large frequencies) has the χ^2 distribution with 1 degree of freedom.

The p value, designated on the basis of the test statistic, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

The χ^2 test of homogeneity for the RR

The Chi-square test of homogeneity for the RR is used in the hypothesis verification that the variable creating stratas, is the modifying effect, i.e. it influences on the designated relative risk in the manner that, the relative risks are significant different for individual stratas.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : & RR_{MH} = RR^{(s)}, \text{ for all the stratas } s = 1, 2, \dots, w, \\ \mathcal{H}_1 : & RR_{MH} \neq RR^{(s)}, \text{ for at least one strata.} \end{aligned}$$

The test statistic, using weighted least squares method, is defined by:

$$\chi^2 = \sum_{s=1}^w v^{(s)} \left(\ln(RR^{(s)}) - \ln(RR_{MH}) \right)^2$$

where:

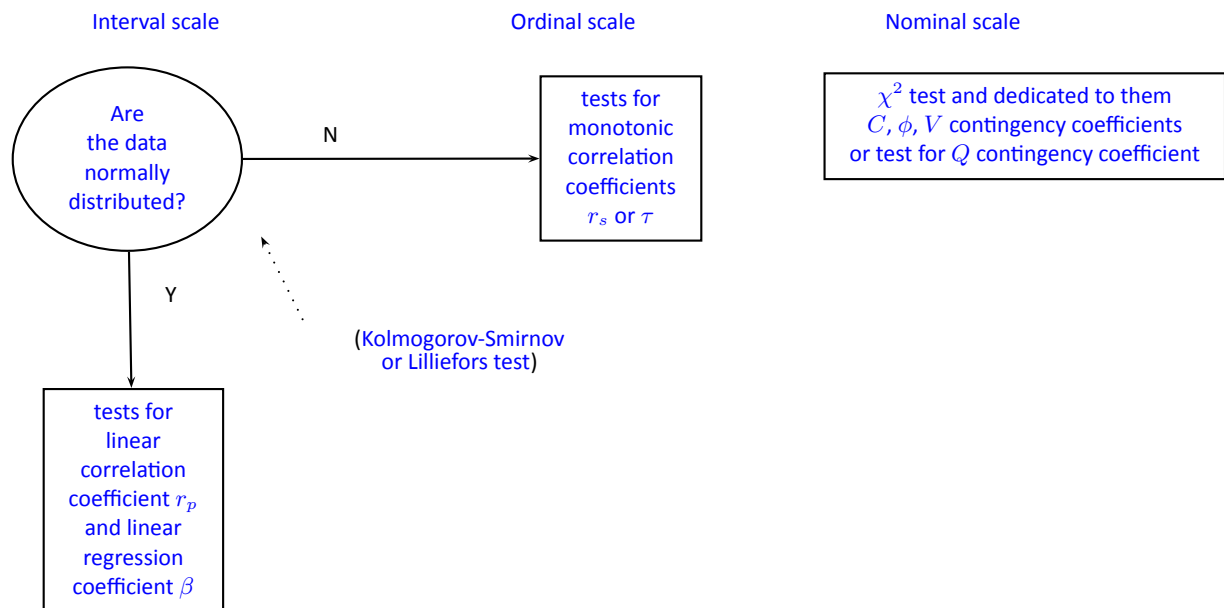
$$v^{(s)} = \left(\frac{O_{12}^{(s)}}{O_{11}^{(s)}(O_{11}^{(s)} + O_{12}^{(s)})} + \frac{O_{22}^{(s)}}{O_{21}^{(s)}(O_{21}^{(s)} + O_{22}^{(s)})} \right)^{-1}.$$

This statistic asymptotically (for large frequencies) has the χ^2 distribution with the number of degrees of freedom calculated using the formula: $df = w - 1$.

The p value, designated on the basis of the test statistic, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

14 CORRELATION



The Correlation coefficients are one of the measures of descriptive statistics which represent the level of correlation (dependence) between 2 or more features (variables). The choice of a particular coefficient depends mainly on the scale, on which the measurements were done. Calculation of coefficients is one of the first steps of the correlation analysis. Then the statistic significance of the gained coefficients may be checked using adequate tests.

Note

Note, that the dependence between variables does not always show the cause-and-effect relationship.

14.1 PARAMETRIC TESTS

14.1.1 THE LINEAR CORRELATION COEFFICIENTS

The **Pearson product-moment correlation coefficient** r_p called also the Pearson's linear correlation coefficient (Pearson (1896,1900)) is used to describe the strength of linear relations between 2 features. It may be calculated on an **interval scale** only if the distribution of the analysed features is a **normal one**.

$$r_p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

where:

x_i, y_i - the following values of the feature X and Y ,

\bar{x}, \bar{y} - means values of features: X and Y ,

n - sample size.

Note

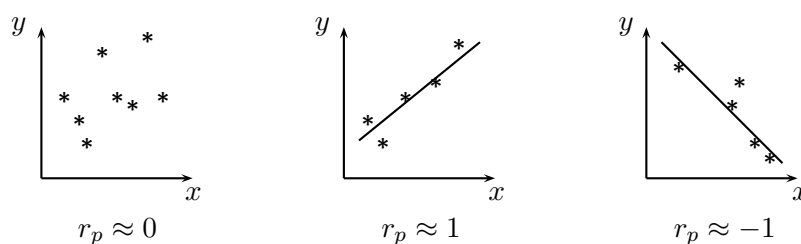
R_p - the Pearson product-moment correlation coefficient in a population;

r_p - the Pearson product-moment correlation coefficient in a sample.

The value of $r_p \in (-1; 1)$, and it should be interpreted the following way:

- $r_p \approx 1$ means a strong positive linear correlation - measurement points are closed to a straight line and when the independent variable increases, the dependent variable increases too;
- $r_p \approx -1$ means a strong negative linear correlation - measurement points are closed to a straight line, but when the independent variable increases, the dependent variable decreases;
- if the correlation coefficient is equal to the value or very closed to zero, there is no linear dependence between the analysed features (but there might exist another relation - a not linear one).

Graph 14.1. Graphic interpretation of r_p .



If one out of the 2 analysed features is constant (it does not matter if the other feature is changed), the features are not dependent from each other. In that situation r_p can not be calculated.

Note

You are not allowed to calculate the correlation coefficient if: there are outliers in a sample (they may make that the value and the sign of the coefficient would be completely wrong), if the sample is clearly heterogeneous, or if the analysed relation takes obviously the other shape than linear.

The coefficient of determination: r_p^2 - reflects the percentage of a dependent variable a variability which is explained by variability of an independent variable.

A created model shows a linear relationship:

$$y = \beta x + \alpha.$$

β and α coefficients of linear regression equation can be calculated using formulas:

$$\beta = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \alpha = \bar{y} - \beta \bar{x}.$$

14.1.2 The test of significance for the Pearson product-moment correlation coefficient

The test of significance for Pearson product-moment correlation coefficient is used to verify the hypothesis determining the lack of linear correlation between an analysed features of a population and it is based on the Pearson's linear correlation coefficient calculated for the sample. The closer to 0 the value of r_p is, the weaker dependence joins the analysed features.

Basic assumptions:

- measurement on the [interval scale](#),
- [normality of distribution](#) of an analysed features in a population.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : R_p &= 0, \\ \mathcal{H}_1 : R_p &\neq 0. \end{aligned}$$

The test statistic is defined by:

$$t = \frac{r_p}{SE},$$

$$\text{where } SE = \sqrt{\frac{1 - r_p^2}{n - 2}}.$$

The value of the test statistic can not be calculated when $r_p = 1$ or $r_p = -1$ or when $n < 3$.

The test statistic has the [t-Student distribution](#) with $n - 2$ degrees of freedom.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned} \text{if } p &\leq \alpha \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p &> \alpha \implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

14.1.3 The test of significance for the coefficient of linear regression equation

This test is used to verify the hypothesis determining the lack of a linear dependence between an analysed features and is based on the slope coefficient (also called an effect), calculated for the sample. The closer to 0 the value of β is, the weaker dependence presents the fitted line.

Basic assumptions:

- measurement on the [interval scale](#),
- [normality of distribution](#) of an analysed features in a population.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : \beta &= 0, \\ \mathcal{H}_1 : \beta &\neq 0.\end{aligned}$$

The test statistic is defined by:

$$t = \frac{\beta}{SE}$$

where:

$$SE = \frac{s_{yx}}{sd_x \sqrt{n-1}},$$

$$s_{yx} = sd_y \sqrt{\frac{n-1}{n-2}(1-r^2)},$$

sd_x, sd_y – standard deviation of the value of features: X and Y .

The value of the test statistic can not be calculated when $r_p = 1$ or $r_p = -1$ or when $n < 3$.

The test statistic has the *t-Student distribution* with $n - 2$ degrees of freedom.

The *p value*, designated on the basis of the *test statistic*, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

Prediction is used to predict the value of a one variable (mainly a dependent variable y_0) on the basis of a value of an another variable (mainly an independent variable x_0). The accuracies of a calculated value are defined by prediction intervals calculated for it.

- **Interpolation** is used to predict the value of a variable, which occurs inside the area for which the regression model was done. Interpolation is mainly a safe procedure - it is assumed only the continuity of the function of analysed variables.
- **Extrapolation** is used to predict the value of variable, which occurs outside the area for which the regression model was done. As opposed to interpolation, extrapolation is often risky and is performed only not far away from the area, where the regression model was created. Similarly to the interpolation, it is assumed the continuity of the function of analysed variables.

The settings window with the Pearson's linear correlation can be opened in Statistics menu → Parametric tests → linear correlation (r-Pearson) or in [Wizard](#).

Linear correlation (r Pearson)

Statistical analysis : Pearson linear correlation

Variable 1

1-age
2-height

Variable 2

1-age
2-height

Test options

☒ Prediction

X value (variable 1) 6

Y value (variable 2)

Data Filter

set of the conditions that are applied to data to produce a subset of your data

All the rules are combined using the logical AND

☒ basic ☐ multiple AND

Report options

☐ Add analysed data

☒ Add graph

0.05 significance level

OK Close

EXAMPLE 14.1. (age-height.pqs file)

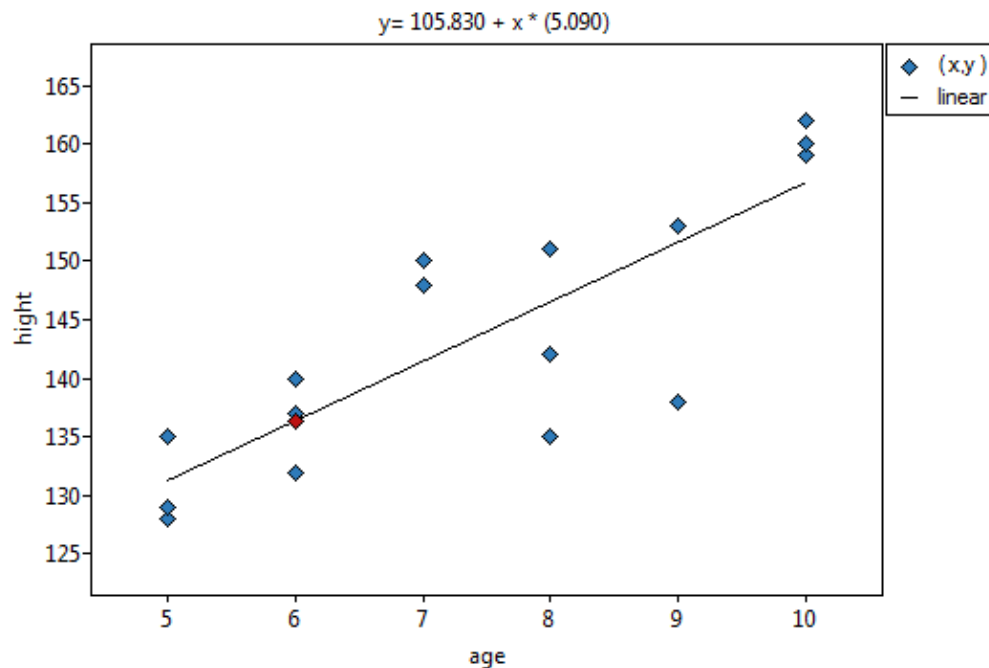
Among some students of a ballet school, the dependence between age and height was analysed. The sample consists of 16 children and the following results of these features (related to the children) were written down:

(age, height): (5, 128) (5, 129) (5, 135) (6, 132) (6, 137) (6, 140) (7, 148) (7, 150) (8, 135) (8, 142) (8, 151) (9, 138) (9, 153) (10, 159) (10, 160) (10, 162).

Hypotheses:

- \mathcal{H}_0 : there is no linear dependence between age and height
for the population of children who attend to the analysed school,
- \mathcal{H}_1 : there is a linear dependence between age and height
for the population of children who attend to the analysed school.

Pearson linear correlation	
Analysis time	0.02sec.
Analysed variables	age,height
Significance level	0.05
Size = number of pairs	16
Group name	age
Group mean	7.4375
Group standard deviation	1.8246
Group name	height
Group mean	143.6875
Group standard deviation	11.187605
The standard deviation of the residuals	6.456417
r	0.830153
r ²	0.689154
Std. err. of r	0.149008
-95% CI for r coefficient	0.568316
+95% CI for r coefficient	0.939318
t-statistic for r	5.571206
Degrees of freedom	14
p-value	0.000069
a - slope	5.090113
Std. err. of a	0.913646
-95% CI for a coefficient	3.130536
+95% CI for a coefficient	7.049689
t-test statistic for a	5.571206
Degrees of freedom	14
p-value	0.000069
b -Y intercept	105.829787
Std. err. of b	6.984318
-95% CI for b coefficient	90.849915
+95% CI for b coefficient	120.809659
t-test statistic for b	15.152487
Degrees of freedom	14
p-value	<0.000001
prediction of Y value for X= 6	136.370463
-95% CI for the prediction of Y	121.821348
+95% CI for the prediction of Y	150.919578



Comparing the p value = 0.000069 with the significance level $\alpha = 0.05$, we draw the conclusion, that there is a linear dependence between age and height in the population of children attending to the analysed school. This dependence is directly proportional, it means that the children grow up as they are getting older.

The Pearson product-moment correlation coefficient, so the strength of the linear relation between age and height counts to $r_p=0.8302$. Coefficient of determination $r_p^2 = 0.6892$ means that about 69% variability of height is explained by the changing of age.

From the regression equation:

$$height = 5.09 \cdot age + 105.83$$

it is possible to calculate the predicted value for a child, for example: in the age of 6. The predicted height of such child is 136.37cm.

14.1.4 The test for checking the equality of the Pearson product-moment correlation coefficients, which come from 2 independent populations

This test is used to verify the hypothesis determining the equality of 2 [Pearson's linear correlation coefficients](#) (R_{p_1}, R_{p_2}).

Basic assumptions:

- r_{p_1} and r_{p_2} come from 2 samples which are chosen randomly from [independent](#) populations,
- r_{p_1} and r_{p_2} describe the strength of dependence of the same features: X and Y ,
- sizes of both samples (n_1 and n_2) are known.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : & R_{p_1} = R_{p_2}, \\ \mathcal{H}_1 : & R_{p_1} \neq R_{p_2}. \end{aligned}$$

The test statistic is defined by:

$$t = \frac{z_{r_{p1}} - z_{r_{p2}}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}},$$

where:

$$z_{r_{p1}} = \frac{1}{2} \ln \left(\frac{1 + r_{p1}}{1 - r_{p1}} \right),$$

$$z_{r_{p2}} = \frac{1}{2} \ln \left(\frac{1 + r_{p2}}{1 - r_{p2}} \right).$$

The test statistic has the *t*-Student distribution with $n_1 + n_2 - 4$ degrees of freedom.

The *p* value, designated on the basis of the test statistic, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

14.1.5 The test for checking the equality of the coefficients of linear regression equation, which come from 2 independent populations

This test is used to verify the hypothesis determining the equality of 2 coefficients of the linear regression equation β_1 and β_2 in analysed populations.

Basic assumptions:

- β_1 and β_2 come from 2 samples which are chosen randomly from independent populations,
- β_1 and β_2 describe the strength of dependence of the same features: X and Y ,
- both sample sizes (n_1 and n_2) are known,
- standard deviations for the values of both features in both samples (sd_{x_1}, sd_{y_1} and sd_{x_2}, sd_{y_2}) are known,
- the Pearson product-moment correlation coefficients of both samples (r_{p1} and r_{p2}) are known.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : \beta_1 &= \beta_2, \\ \mathcal{H}_1 : \beta_1 &\neq \beta_2. \end{aligned}$$

The test statistic is defined by:

$$t = \frac{\beta_1 - \beta_2}{\sqrt{\frac{s_{yx_1}^2}{sd_{x_1}^2(n_1-1)} + \frac{s_{yx_2}^2}{sd_{x_2}^2(n_2-1)}}},$$

where:

$$s_{yx_1} = sd_{y_1} \sqrt{\frac{n_1 - 1}{n_1 - 2} (1 - r_{p1}^2)},$$

$$s_{yx_2} = sd_{y_2} \sqrt{\frac{n_2 - 1}{n_2 - 2} (1 - r_{p2}^2)}.$$

The test statistic has the *t*-Student distribution with $n_1 + n_2 - 4$ degrees of freedom.

The *p* value, designated on the basis of the test statistic, is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

The settings window with the comparison of correlation coefficients can be opened in Statistics menu \rightarrow Parametric tests \rightarrow comparison of correlation coefficients.

Comparison of correlation coefficients

Statistical analysis : Comparison of the correlation coefficients

fill with saved selection ▼

	G1	G2
slope a	2	3
Pearson correlation coefficient	0.6	0.67
sample size n	60	72
Std. dev. for X	0.441	0.556
Std. dev. for Y	0.575	0.239

Report options

☒ Add analysed data

0.05 ▼ significance level

OK Close

14.2 NONPARAMETRIC TESTS

14.2.1 THE MONOTONIC CORRELATION COEFFICIENTS

The monotonic correlation may be described as monotonically increasing or monotonically decreasing. The relation between 2 features is presented by the monotonic increasing if the increasing of the one feature accompanies with the increasing of the other one. The relation between 2 features is presented by the monotonic decreasing if the increasing of the one feature accompanies with the decreasing of the other one.

The **Spearman's rank-order correlation coefficient** r_s is used to describe the strength of monotonic relations between 2 features: X and Y . It may be calculated on an **ordinal** scale or an **interval** one. The value of the Spearman's rank correlation coefficient should be calculated using the following formula:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

where:

$d_i = R_{x_i} - R_{y_i}$ – difference of **rank**s for the feature X and Y ,
 n number of d_i .

This formula is modified when there are **ties**:

$$r_s = \frac{\Sigma_X + \Sigma_Y - \sum_{i=1}^n d_i^2}{2\sqrt{\Sigma_X \Sigma_Y}},$$

where:

$$\Sigma_X = \frac{n^3 - n - T_X}{12}, \Sigma_Y = \frac{n^3 - n - T_Y}{12},$$

$$T_X = \sum_{i=1}^s (t_{i(X)}^3 - t_{i(X)}), T_Y = \sum_{i=1}^s (t_{i(Y)}^3 - t_{i(Y)}),$$

t – number of cases included in tie.

This correction is used, when ties occur. If there are no ties, the correction is not calculated, because the correction is reduced to the formula describing the above equation.

Note

R_s – the Spearman's rank correlation coefficient in a population;

r_s – the Spearman's rank correlation coefficient in a sample.

The value of $r_s \in (-1; 1)$, and it should be interpreted the following way:

- $r_s \approx 1$ means a strong positive monotonic correlation (increasing) – when the independent variable increases, the dependent variable increases too;
- $r_s \approx -1$ means a strong negative monotonic correlation (decreasing) – when the independent variable increases, the dependent variable decreases;
- if the Spearman's correlation coefficient is of the value equal or very close to zero, there is no monotonic dependence between the analysed features (but there might exist another relation - a non monotonic one, for example a sinusoidal relation).

The **Kendall's $\tilde{\tau}$ correlation coefficient** (Kendall (1938)[42]) is used to describe the strength of monotonic relations between features. It may be calculated on an **ordinal** scale or **interval** one. The value of the Kendall's $\tilde{\tau}$ correlation coefficient should be calculated using the following formula:

$$\tilde{\tau} = \frac{2(n_C - n_D)}{\sqrt{n(n-1) - T_X} \sqrt{n(n-1) - T_Y}},$$

where:

n_C – number of pairs of observations, for which the values of the **rank**s for the X feature as well as Y feature are changed in the same direction (the number of agreed pairs),

n_D – number of pairs of observations, for which the values of the ranks for the X feature are changed in the different direction than for the Y feature (the number of disagreed pairs),

$$T_X = \sum_{i=1}^s (t_{i(X)}^2 - t_{i(X)}), T_Y = \sum_{i=1}^s (t_{i(Y)}^2 - t_{i(Y)}),$$

t – number of cases included in a tie.

The formula for the $\tilde{\tau}$ correlation coefficient includes the correction for ties. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $T_X = 0$ i $T_Y = 0$).

Note

τ – the Kendall's correlation coefficient in a population;

$\tilde{\tau}$ – the Kendall's correlation coefficient in a sample.

The value of $\tilde{\tau} \in < -1; 1 >$, and it should be interpreted the following way:

- $\tilde{\tau} \approx 1$ means a strong agreement of the sequence of ranks (the increasing monotonic correlation) – when the independent variable increases, the dependent variable increases too;
- $\tilde{\tau} \approx -1$ means a strong disagreement of the sequence of ranks (the decreasing monotonic correlation) – when the independent variable increases, the dependent variable decreases;
- if the Kendall's $\tilde{\tau}$ correlation coefficient is of the value equal or very close to zero, there is no monotonic dependence between analysed features (but there might exist another relation - a non monotonic one, for example a sinusoidal relation).

The Spearman's r_s versus the Kendall's $\tilde{\tau}$

- for an interval scale with a normality of the distribution, the r_s gives the results which are close to r_p , but $\tilde{\tau}$ may be totally different from r_p ,
- the $\tilde{\tau}$ value is less or equal to r_p value,
- the $\tilde{\tau}$ is an unbiased estimator of the population parameter τ , while the r_s is a biased estimator of the population parameter R_s .

14.2.2 The test of significance for the Spearman's rank-order correlation coefficient

The test of significance for the Spearman's rank-order correlation coefficient is used to verify the hypothesis determining the lack of monotonic correlation between analysed features of the population and it is based on the Spearman's rank-order correlation coefficient calculated for the sample. The closer to 0 the value of r_s is, the weaker dependence joins the analysed features.

Basic assumptions:

- measurement on an **ordinal scale** or on an **interval scale**.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : R_s &= 0, \\ \mathcal{H}_1 : R_s &\neq 0. \end{aligned}$$

The test statistic is defined by:

$$t = \frac{r_s}{SE},$$

where $SE = \sqrt{\frac{1 - r_s^2}{n - 2}}$.

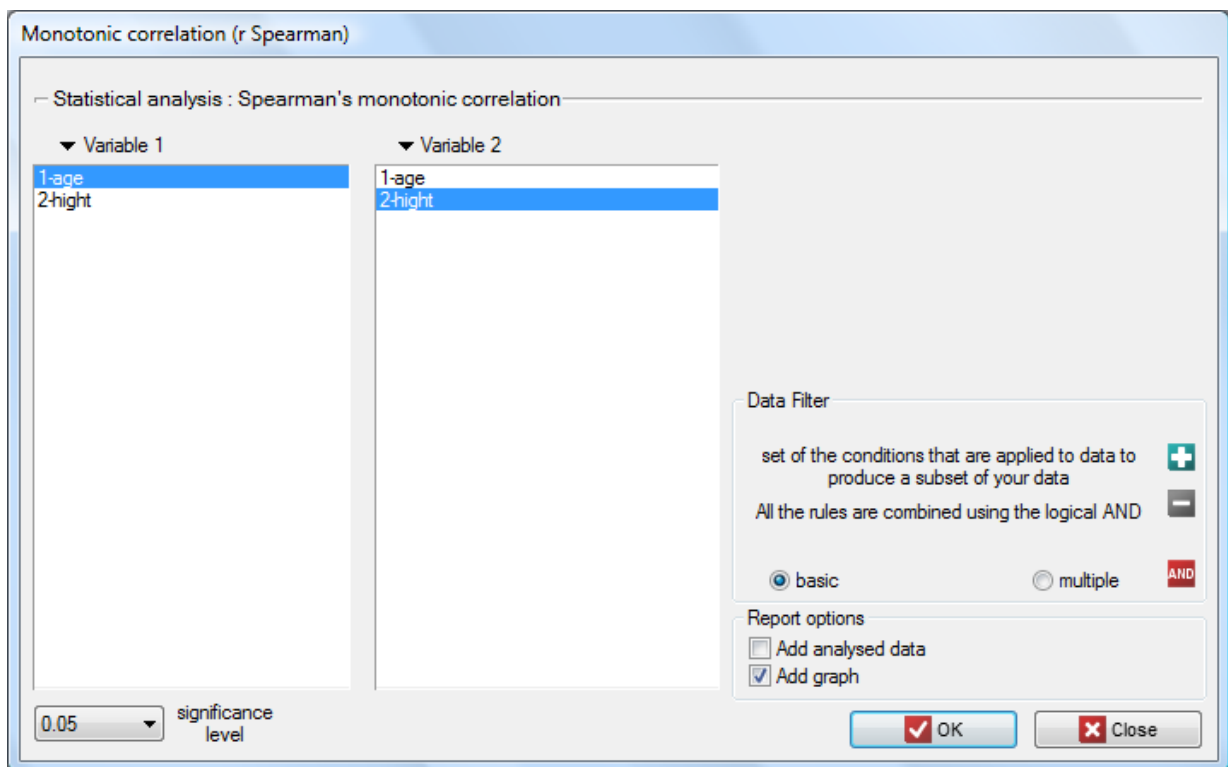
The value of the test statistic can not be calculated when $r_s = 1$ lub $r_s = -1$ or when $n < 3$.

The test statistic has the *t-Student distribution* with $n - 2$ degrees of freedom.

The *p value*, designated on the basis of the *test statistic*, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

The settings window with the Spearman's monotonic correlation can be opened in Statistics menu → NonParametric tests (ordered categories) → monotonic correlation (r-Spearman) or in [Wizard](#).

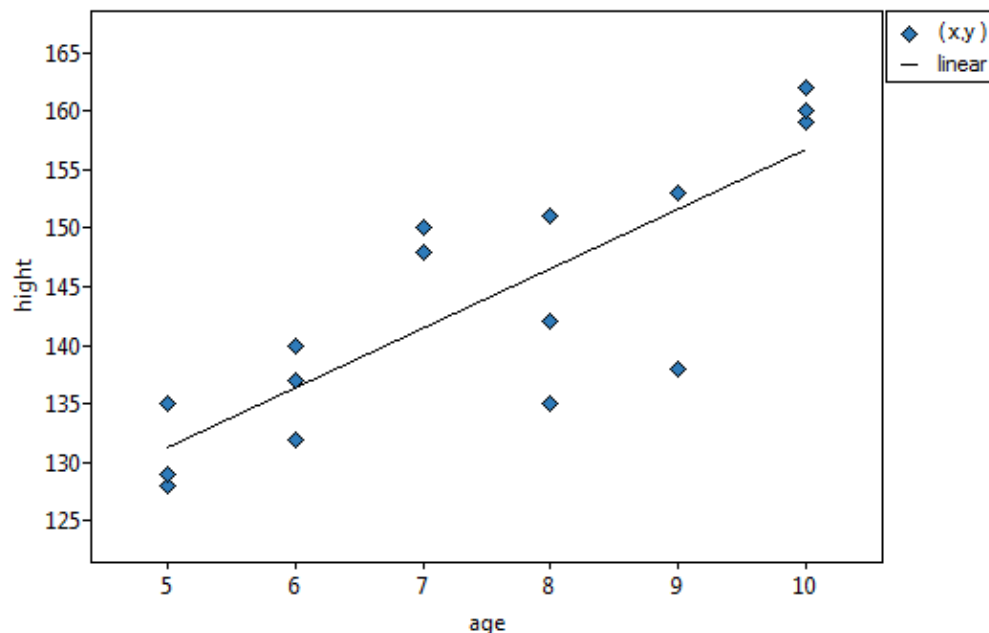


EXAMPLE (14.1) continuation (*age-height.pqs file*)

Hypotheses:

- \mathcal{H}_0 : there is no monotonic dependence between age and height for the population of children attending to the analysed school,
- \mathcal{H}_1 : there is a monotonic dependence between age and height for the population of children attending to the analysed school.

Spearman's monotonic correlation	
Analysis time	0.03sec.
Analysed variables	age,height
Significance level	0.05
Size = number of pairs	16
r	0.839739
Std. err. of r	0.14512
-95% CI for r coefficient	0.578742
+95% CI for r coefficient	0.944696
t-statistic for r	5.786513
Degrees of freedom	14
p-value	0.000047



Comparing the p value = 0.000047 with the significance level $\alpha = 0.05$, we draw the conclusion that there is a monotonic dependence between age and height in the population of children attending to the analysed school. This dependence is directly proportional, it means that children grow up as they get older. The Spearman's rank correlation coefficient, so the strength of a monotonic relation between age and height counts to $r_s=0.8397$.

14.2.3 The test of significance for the Kendall's tau correlation coefficient

The test of significance for the Kendall's $\tilde{\tau}$ correlation coefficient is used to verify the hypothesis determining the lack of monotonic correlation between analysed features of population. It is based on the Kendall's tau correlation coefficient calculated for the sample. The closer to 0 the value of $\tilde{\tau}$ is, the weaker dependence joins the analysed features.

Basic assumptions:

- measurement on an **ordinal scale** or on an **interval scale**.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : \tau &= 0, \\ \mathcal{H}_1 : \tau &\neq 0.\end{aligned}$$

The test statistic is defined by:

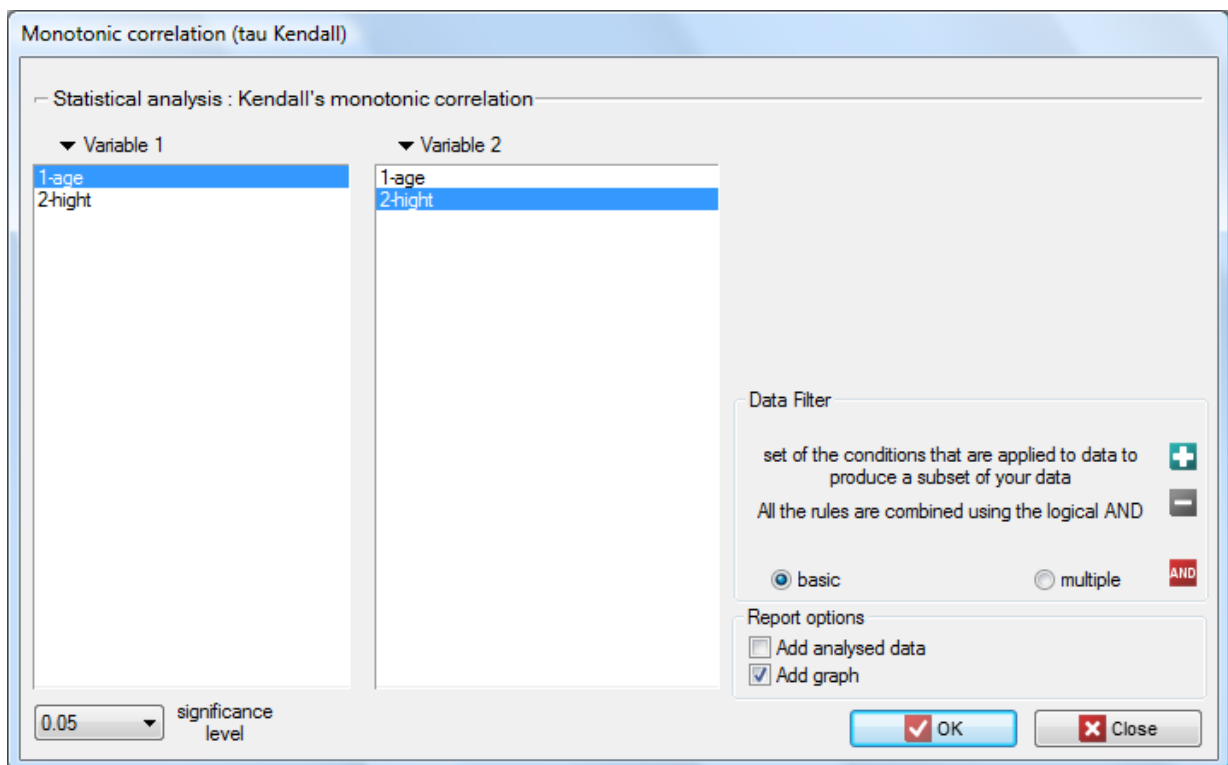
$$Z = \frac{3\tilde{\tau}\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}.$$

The test statistic asymptotically (for a large sample size) has the **normal distribution**.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The settings window with the Kendall's monotonic correlation can be opened in Statistics menu → NonParametric tests (ordered categories) → monotonic correlation (tau-Kendall) or in **Wizard**.

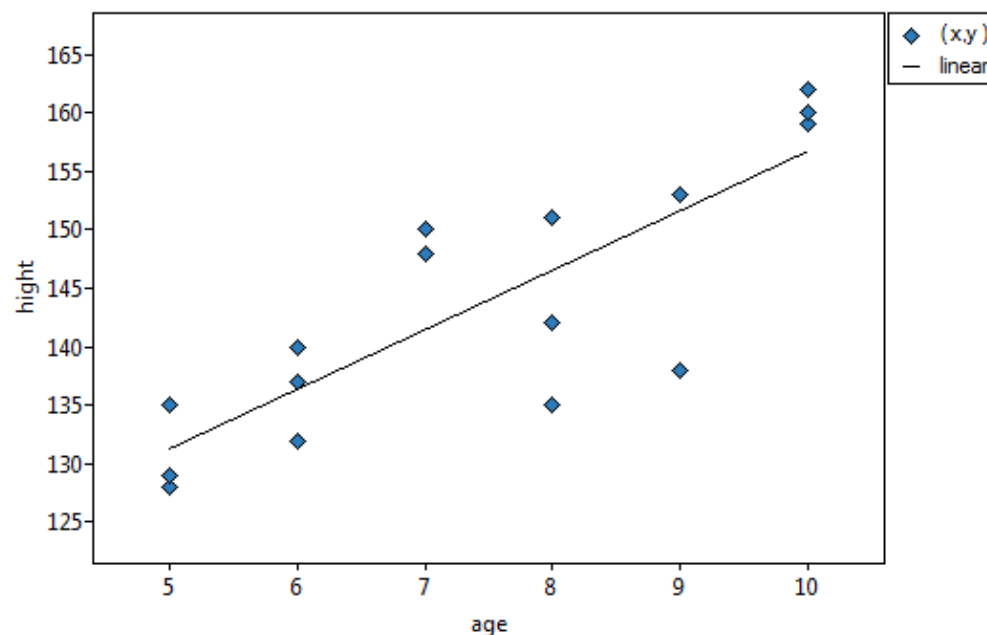


EXAMPLE (14.1) continuation (*age-height.pqs file*)

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : & \text{there is no monotonic dependence between age and height} \\ & \text{for the population of children attending to the analysed school,} \\ \mathcal{H}_1 : & \text{there is a monotonic dependence between age and height} \\ & \text{for the population of children attending to the analysed school.}\end{aligned}$$

Kendall's monotonic correlation	
Analysis time	0.03sec.
Analysed variables	age,height
Significance level	0.05
Size = number of pairs	16
tau	0.721205
Z statistic for tau	3.896455
p-value (asymptotic)	0.000098



Comparing the p value = 000098 with the significance level $\alpha = 0.05$, we draw the conclusion that there is a monotonic dependence between age and height in the population of children attending to the analysed school. This dependence is directly proportional, it means that children grow up as they get older. The Kendall's correlation coefficient, so the strength of a monotonic relation between age and height counts to $\tilde{r}=0.7212$.

14.2.4 CONTINGENCY TABLES COEFFICIENTS AND THEIR STATISTICAL SIGNIFICANCE

The contingency coefficients are calculated for the [raw data](#) or the data gathered in a [contingency table](#) (look at the table (11.1)).

The Yule's Q contingency coefficient

The Yule's Q contingency coefficient (Yule, 1900[88]) is a measure of correlation, which can be calculated for 2×2 contingency tables.

$$Q = \frac{O_{11}O_{22} - O_{12}O_{21}}{O_{11}O_{22} + O_{12}O_{21}},$$

where:

$O_{11}, O_{12}, O_{21}, O_{22}$ - observed frequencies in a [contingency table](#).

The Q coefficient value is included in a range of $< -1; 1 >$. The closer to 0 the value of the Q is, the weaker dependence joins the analysed features, and the closer to -1 or $+1$, the stronger dependence joins the analysed features. There is one disadvantage of this coefficient. It is not much resistant to small observed frequencies (if one of them is 0, the coefficient might wrongly indicate the total dependence of features).

The statistic significance of the Yule's Q coefficient is defined by the Z test.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : Q &= 0, \\ \mathcal{H}_1 : Q &\neq 0.\end{aligned}$$

The test statistic is defined by:

$$Z = \frac{Q}{\sqrt{\frac{1}{4}(1 - Q^2)^2 \left(\frac{1}{O_{11}} + \frac{1}{O_{12}} + \frac{1}{O_{21}} + \frac{1}{O_{22}} \right)}}.$$

The test statistic asymptotically (for a large sample size) has the [normal distribution](#).

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The ϕ contingency coefficient

The Phi contingency coefficient is a measure of correlation, which can be calculated for 2×2 contingency tables.

$$\phi = \sqrt{\frac{\chi^2}{n}},$$

where:

χ^2 – value of the χ^2 test statistic,

n – total frequency in a [contingency table](#).

The ϕ coefficient value is included in a range of $< 0; 1 >$. The closer to 0 the value of ϕ is, the weaker dependence joins the analysed features, and the closer to 1, the stronger dependence joins the analysed features.

The ϕ contingency coefficient is considered as **statistically significant**, if the [p-value](#) calculated on the basis of the χ^2 test (designated for this table) is equal to or less than the significance level α .

The settings window with the measures of correlation Q-Yule, Phi can be opened in Statistics menu \rightarrow NonParametric tests (unordered categories) \rightarrow Q-Yule, Phi (2x2) or in [Wizard](#).

Q-Yule, Phi (2x2)

Statistical analysis : Measures of the correlation Q-Yule, Phi

Data reduced by the selected area [2x2]

fill with saved selection

	A	B
1	50	40
2	20	60

☒ Contingency table
 ☐ Raw data

Report options

☐ Add analysed data
☒ Add graph
☐ Add percentages

Rows

0.05 significance level

OK Close

The Cramer's V contingency coefficient

The Cramer's V contingency coefficient (Cramer, 1946[24]), is an extension of the ϕ coefficient on $r \times c$ contingency tables.

$$V = \sqrt{\frac{\chi^2}{n(w-1)}},$$

where:

- χ^2 – value of the χ^2 test statistic,
- n – total frequency in a contingency table,
- w – the smaller the value out of r and c .

The V coefficient value is included in a range of $< 0; 1 >$. The closer to 0 the value of V is, the weaker dependence joins the analysed features, and the closer to 1, the stronger dependence joins the analysed features. The V coefficient value depends also on the table size, so you should not use this coefficient to compare different sizes of contingency tables.

The V contingency coefficient is considered as **statistically significant**, if the p -value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than the significance level α .

The Pearson's C contingency coefficient

The Pearson's C contingency coefficient is a measure of correlation, which can be calculated for $r \times c$ contingency tables.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

where:

- χ^2 – value of the χ^2 test statistic,
- n – total frequency in a contingency table.

The C coefficient value is included in a range of $< 0; 1$). The closer to 0 the value of C is, the weaker dependence joins the analysed features, and the farther from 0, the stronger dependence joins the analysed features. The C coefficient value depends also on the table size (the bigger table, the closer to 1 C value can be), that is why it should be calculated the top limit, which the C coefficient may gain – for the particular table size:

$$C_{max} = \sqrt{\frac{w-1}{w}},$$

where:

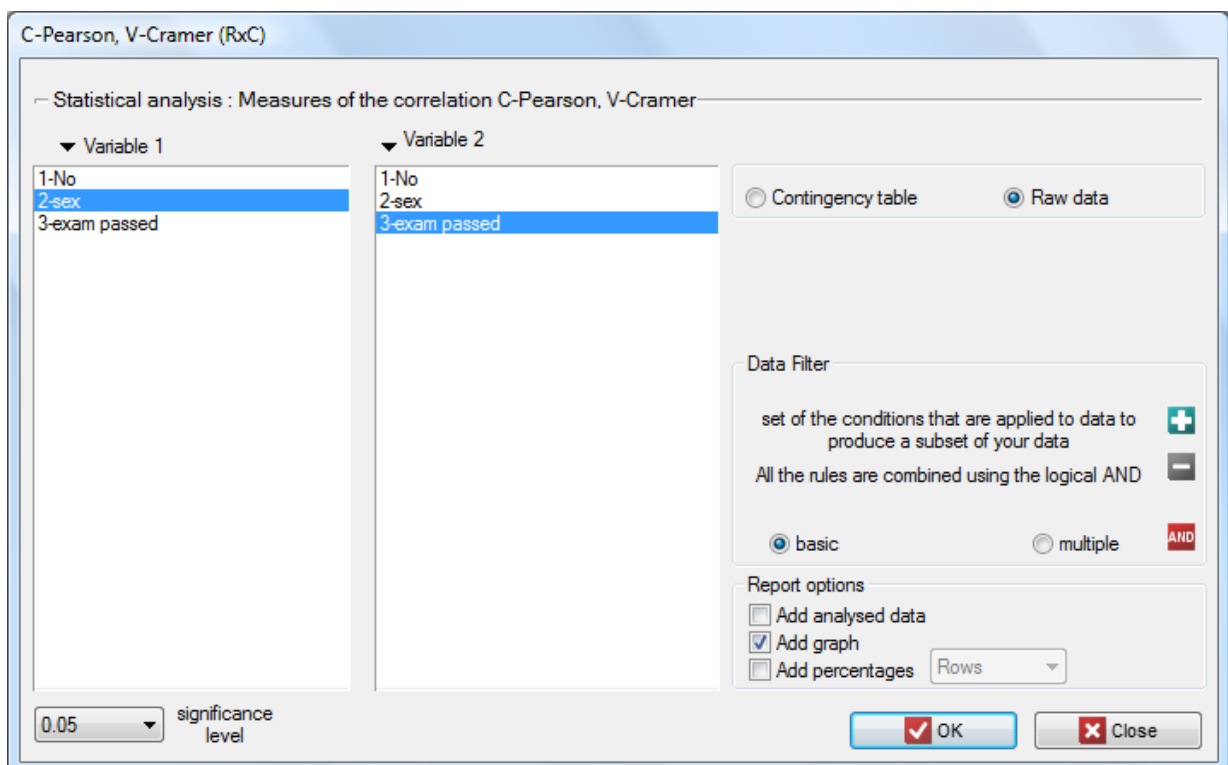
w – the smaller value out of r and c .

An uncomfortable consequence of dependence of C value on a table size is the lack of possibility of comparison the C coefficient value calculated for the various sizes of contingency tables. A little bit better measure is a contingency coefficient adjusted for the table size (C_{adj}):

$$C_{adj} = \frac{C}{C_{max}}.$$

The C contingency coefficient is considered as **statistically significant**, if the p -value calculated on the basis of the χ^2 test (designated for this table) is equal to or less than significance level α .

The settings window with the measures of correlation C-Pearson, V-Cramer can be opened in Statistics menu → NonParametric tests (unordered categories) → C-Pearsona, V-Cramera (RxC) or in [Wizard](#).



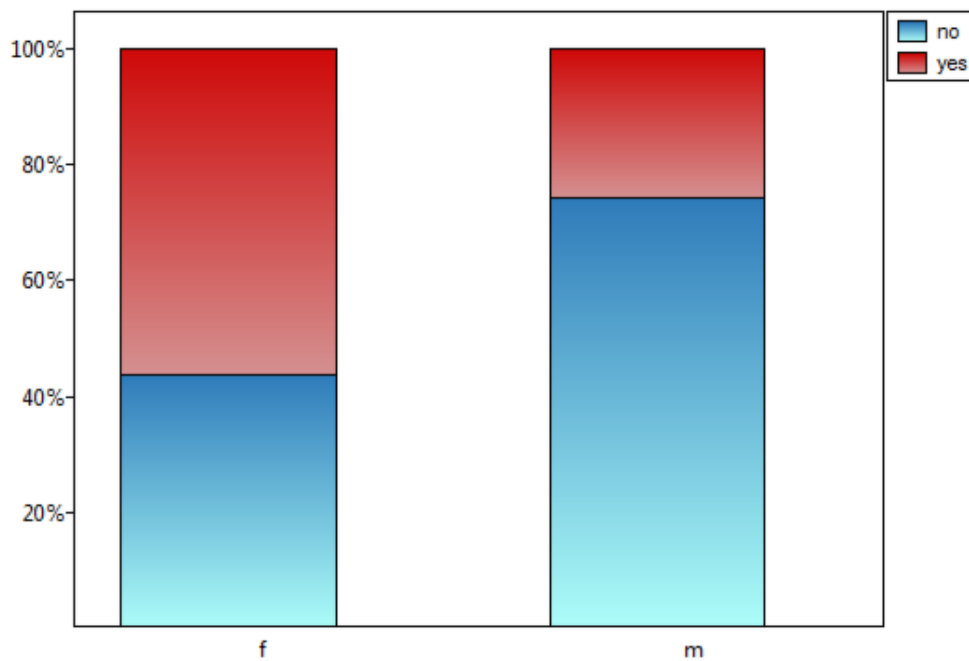
EXAMPLE 14.2. (sex-exam.pqs file)

There is a sample of 170 persons ($n = 170$), who have 2 features analysed (X =sex, Y =passing the exam). Each of these features occurs in 2 categories ($X_1=f$, $X_2=m$, $Y_1=yes$, $Y_2=no$). Basing on the sample, we would like to get to know, if there is any dependence between sex and passing the exam in an analysed population. The data distribution is presented in a contingency table:

Observed frequencies O_{ij}		passing the exam		
		yes	no	total
sex	f	50	40	90
	m	20	60	80
	total	70	100	170

Measures of the correlation Q-Yule, Phi	
Analysis time	0.02sec.
Analysed variables	Contingency table
Significance level	0.05
Size	170
Phi	0.30989
Chi-square statistic	16.325397
Degrees of freedom	1
p-value	0.000053
Q-Yule	0.578947
Z statistic	5.211986
p-value (asymptotic)	<0.000001

Measures of the correlation C-Pearson, V-Cramer	
Analysis time	0.03sec.
Analysed variables	sex;exam passed
Significance level	0.05
Size	170
C-Pearson	0.296003
C-Pearson (max)	0.707107
C-Pearson (adjusted)	0.418611
V-Cramer	0.30989
Chi-square statistic	16.325397
Degrees of freedom	1
p-value	0.000053



The test statistic value is $\chi^2 = 16.33$ and the p value calculated for it: $p = 0.00005$. The result indicates that there is a statistically significant dependence between sex and passing the exam in the analysed population.

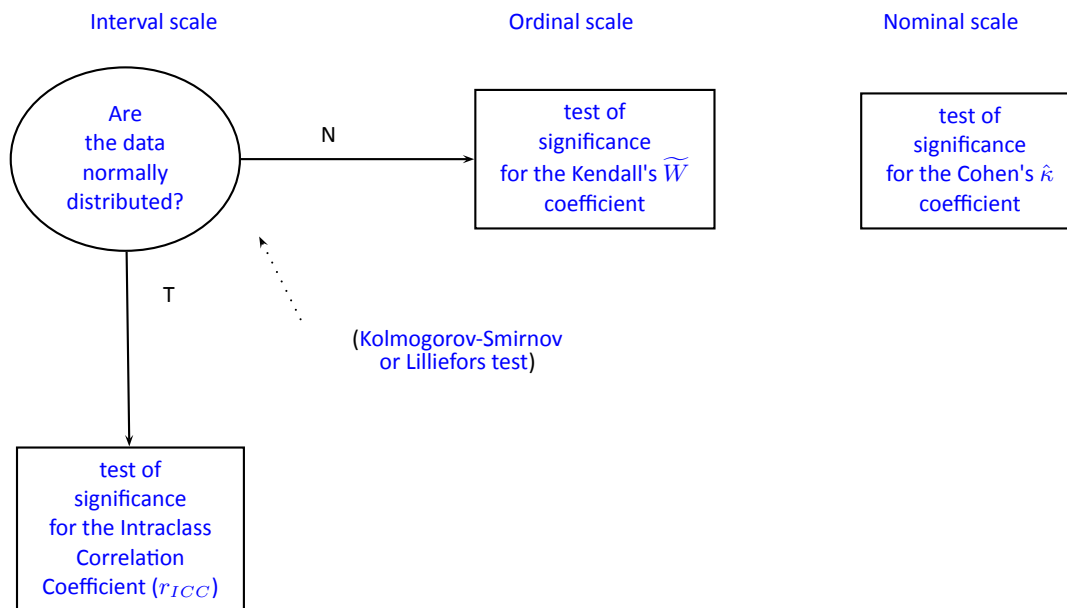
Coefficient values, which are based on the χ^2 test, so the strength of the correlation between analysed features are:

$$C_{adj}\text{-Pearson} = 0.42.$$

$$V\text{-Cramer} = \phi = 0.31$$

The $Q\text{-Yule} = 0.58$, and the p value of the Z test (similarly to χ^2 test) indicates the statistically significant dependence between the analysed features.

15 AGREEMENT ANALYSIS



15.1 PARAMETRIC TESTS

15.1.1 The intraclass correlation coefficient and the test of its significance

The intraclass correlation coefficient is used when the measurement of variables is done by a few "judges" ($k \geq 2$). It measures the strength of **interjudge reliability** — the degree of its assessment concordance.

If the distribution of a variable is a **normal distribution**, it can be represented in a **dependent model** for the **interval scale**.

$$r_{ICC} = \frac{MS_{BS} - MS_{res}}{MS_{BS} + (k - 1)MS_{res} + \frac{k}{n}(MS_{BC} - MS_{res})},$$

where:

MS_{BC} — mean square between-conditions (between judges) — check **ANOVA for dependent groups**,

MS_{BS} — mean square between-subjects,

MS_{res} — mean square residual,

n — sample size,

k — number of judges.

Note

R_{ICC} — the intraclass correlation coefficient in a population;

r_{ICC} — the intraclass correlation coefficient in a sample.

The value of $r_{ICC} \in (-1; 1)$ and it should be interpreted in the following way:

- $r_{ICC} \approx 1$ it is an absolute concordance of objects assessment made by judges; it is especially reflected in a high-variance between objects (a significant means difference between n objects) and a low-variance between judges assessments (a small means difference of assessments designated by k judges);
- $r_{ICC} \approx -1$ a negative intraclass coefficient is treated in the same ways as $r_{ICC} \approx 0$;
- $r_{ICC} \approx 0$ a lack of an absolute concordance in individual objects assessments made by judges; it is visible in a small variance between objects (a small means difference between objects) and in a large variance between judges assessments (a significant means difference of assessments designated by k judges).

In addition, an **average intraclass correlation coefficient** can be formulated as:

$$\bar{r}_{ICC} = \frac{k \cdot ICC}{1 + (k - 1)ICC}.$$

If we averaged these two judges assessments and used them as a one result, the coefficient would not be directly related to the problem, but to the reliability of the situation results.

The F test of significance for the intraclass correlation coefficient

Basic assumptions:

- measurement on an **interval scale**,

- the **normal distribution** for all variables which are the differences of measurement pairs (or the normal distribution for an analysed variable in each measurement).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : R_{ICC} &= 0 \\ \mathcal{H}_1 : R_{ICC} &\neq 0 \quad (R_{ICC} = 1)\end{aligned}$$

The test statistic is defined by:

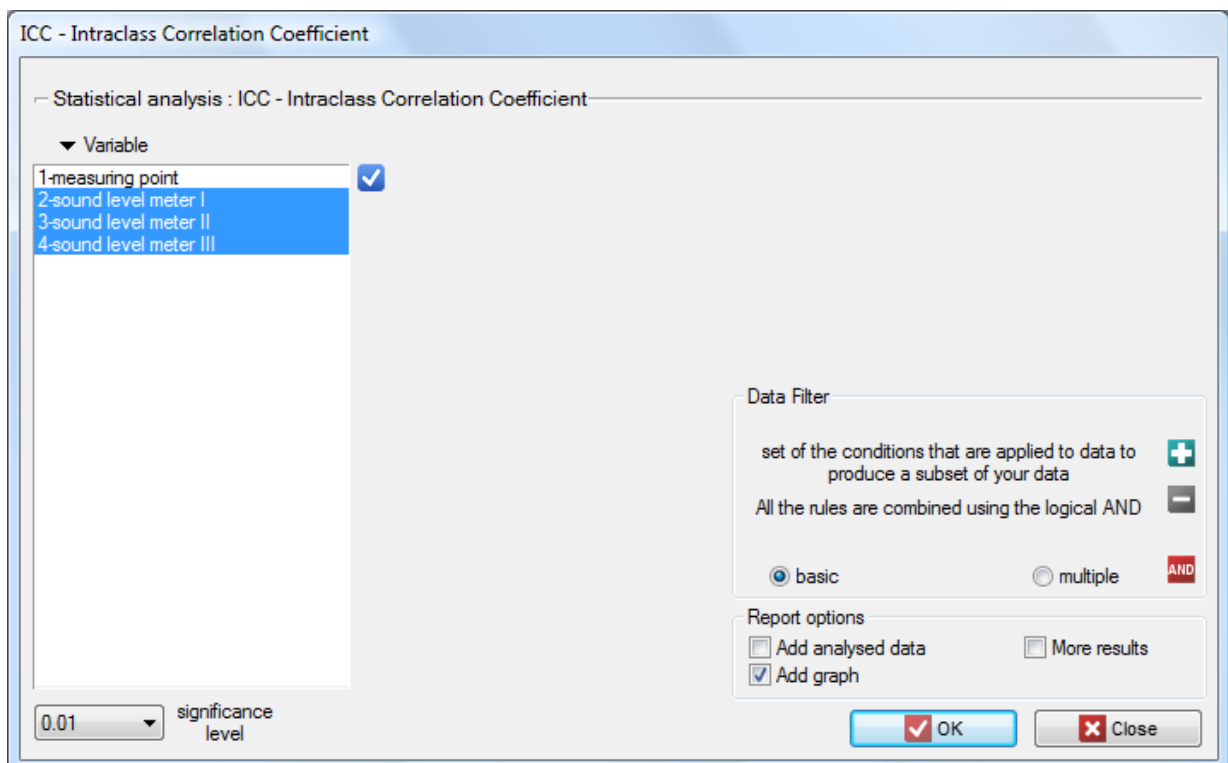
$$F = \frac{MS_{BS}}{MS_{res}}$$

This statistic has the **F Snedecor distribution** with $df_{BS} = n - 1$ and $df_{res} = (n - 1)(k - 1)$ degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The settings window with the ICC – Intraclass Correlation Coefficient can be opened in Statistics menu→Parametric tests→ICC – Intraclass Correlation Coefficient or in [Wizard](#).



EXAMPLE 15.1. (sound intensity.pqs file)

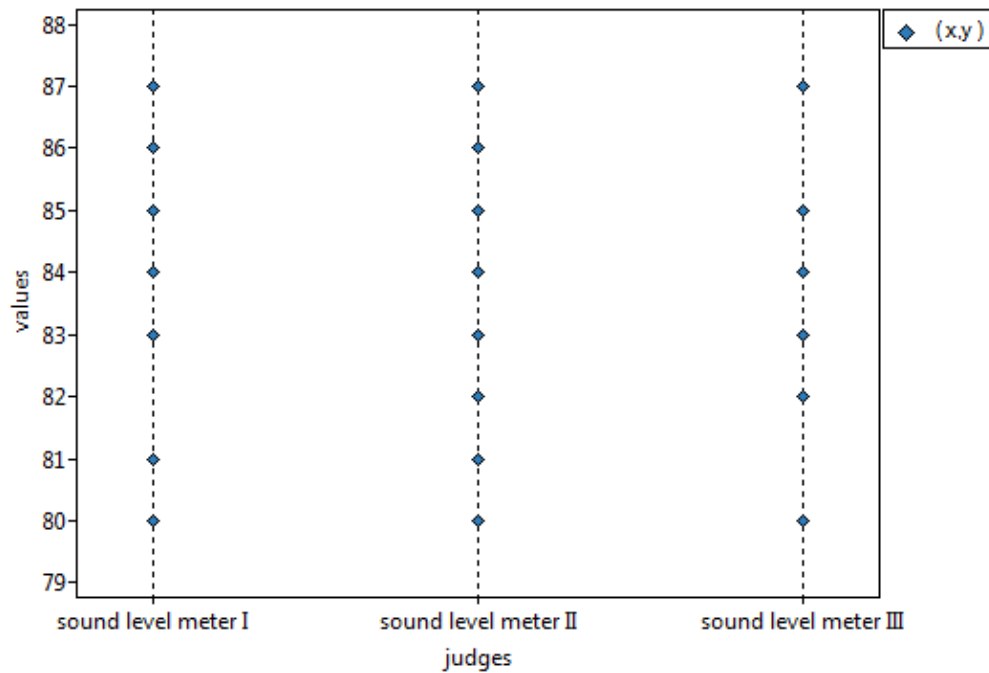
The concordance of sound intensity was measured by three different meters. The measurements were done in 12 different measuring points.

measuring point	meter I	meter II	meter III
A	84	84	84
B	85	85	84
C	84	84	85
D	87	87	87
E	85	86	85
F	80	80	80
G	81	81	82
H	86	86	87
I	83	82	83
J	84	82	84
K	83	82	83
L	84	83	84

Hypotheses:

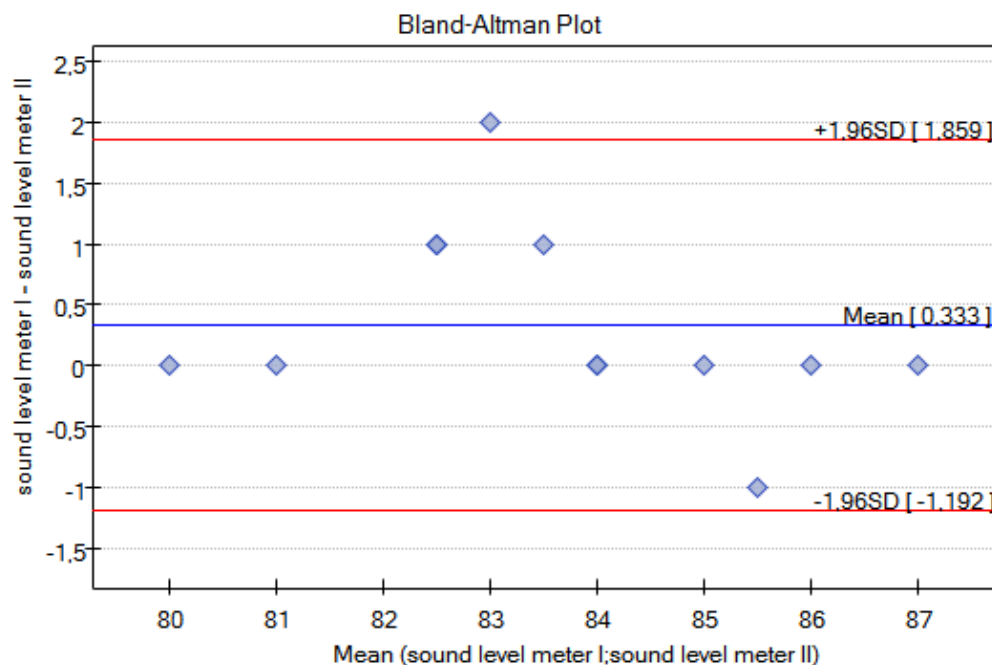
- \mathcal{H}_0 : a lack of an absolute concordance between the levels of sound intensity measured by three different meters, in the population represented by the sample,
- \mathcal{H}_1 : the levels of sound intensity, measured in the population represented by the sample, are absolutely concordant.

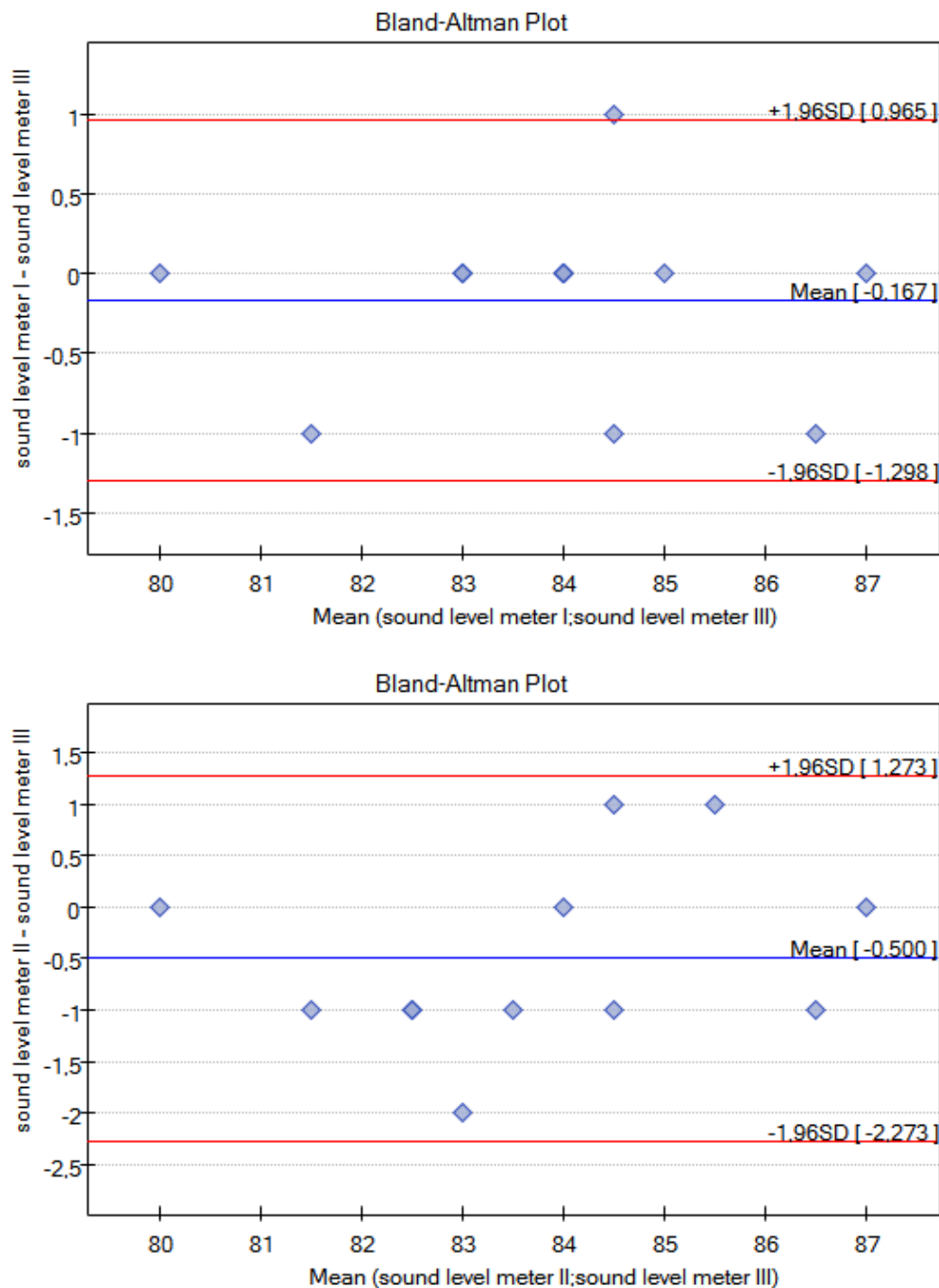
ICC - Intraclass Correlation Coefficient	
Analysis time	0.02sec.
Analysed variables	sound level meter I,sound level meter II,sound level meter III
Significance level	0.05
Total sum of squares (SS[T])	138.222222
Between-conditions sum of squares (SS[BC])	1.555556
Between-subjects sum of squares (SS[BS])	130.222222
Residual sum of squares (SS[RES])	6.444444
Between-conditions degrees of freedom (df[BC])	2
Between-subjects degrees of freedom (df[BS])	11
Residual degrees of freedom (df[RES])	22
Total degrees of freedom (df[T])	35
Mean square between-conditions (MS[BC])	0.777778
Mean square between-subjects (MS[BS])	11.838384
Mean square residual (MS[RES])	0.292929
Intraclass correlation coefficient (r[ICC])	0.92029
Average measure r[ICC]	0.971939
F statistic	40.413793
p-value	<0.000001



Comparing the $p < 0,000001$ with the significance level $\alpha = 0.05$, we have stated that the sound intensity levels, measured by three different meters, are absolutely concordant in the analysed population. The strength of absolute concordance is high: $r_{ICC} = 0.92029$.

Concordance of the results we also see in the Bland-Altman plots[3][10], where almost all of the values fall into the specified range:





15.2 NONPARAMETRIC TESTS

15.2.1 The Kendall's coefficient of concordance and the test of its significance

The Kendall's \widetilde{W} coefficient of concordance is described in the works of Kendall, Babington-Smith (1939)[43] and Wallis (1939)[80]. It is used when the result comes from different sources (from different judges) and concerns a few ($k \geq 2$) objects. However, the assessment concordance is necessary. Is often used in measuring the **interjudge reliability** strength – the degree of (judges) assessment concordance.

The Kendall's coefficient of concordance is calculated on an [ordinal scale](#) or a [interval scale](#). Its value is

calculated according to the following formula:

$$\widetilde{W} = \frac{12U - 3n^2k(k+1)^2}{n^2k(k^2 - 1) - nC},$$

where:

n – number of different assessments sets (the number of judges),

k – number of ranked objects,

$$U = \sum_{j=1}^k \left(\sum_{i=1}^n R_{ij} \right)^2,$$

R_{ij} – ranks ascribed to the following objects ($j = 1, 2, \dots, k$), independently for each judge ($i = 1, 2, \dots, n$),

$C = \sum (t^3 - t)$ – a correction for ties,

t – number of cases incorporated into tie.

The coefficient's formula includes C – the correction for ties. This correction is used, when ties occur (if there are no ties, the correction is not calculated, because of $C = 0$).

Note

\widetilde{W} – the Kendall's coefficient in a population;

\widehat{W} – the Kendall's coefficient in a sample.

The value of $\widetilde{W} \in [-1; 1]$ and it should be interpreted in the following way:

- $\widetilde{W} \approx 1$ means a strong concordance in judges assessments;
- $\widetilde{W} \approx 0$ means a lack of concordance in judges assessments.

The Kendall's \widetilde{W} coefficient of concordance vs. the Spearman r_s coefficient:

When the values of the Spearman r_s correlation coefficient (for all possible pairs) are calculated, the **average r_s coefficient** – marked by \bar{r}_s is a linear function of \widetilde{W} coefficient:

$$\bar{r}_s = \frac{n\widetilde{W} - 1}{n - 1}$$

The Kendall's \widetilde{W} coefficient of concordance vs. the Friedman ANOVA:

The Kendall's \widetilde{W} coefficient of concordance and the **Friedman ANOVA** are based on the same mathematical model. As a result, the value of the chi-square test statistic for the Kendall's coefficient of concordance and the value of the chi-square test statistic for the Friedman ANOVA are the same.

The chi-square test of significance for the Kendall's coefficient of concordance

Basic assumptions:

- measurement on an **ordinal scale** or on an **interval scale**.

Hypotheses:

$$\mathcal{H}_0 : W = 0$$

$$\mathcal{H}_1 : W \neq 0$$

The test statistic is defined by:

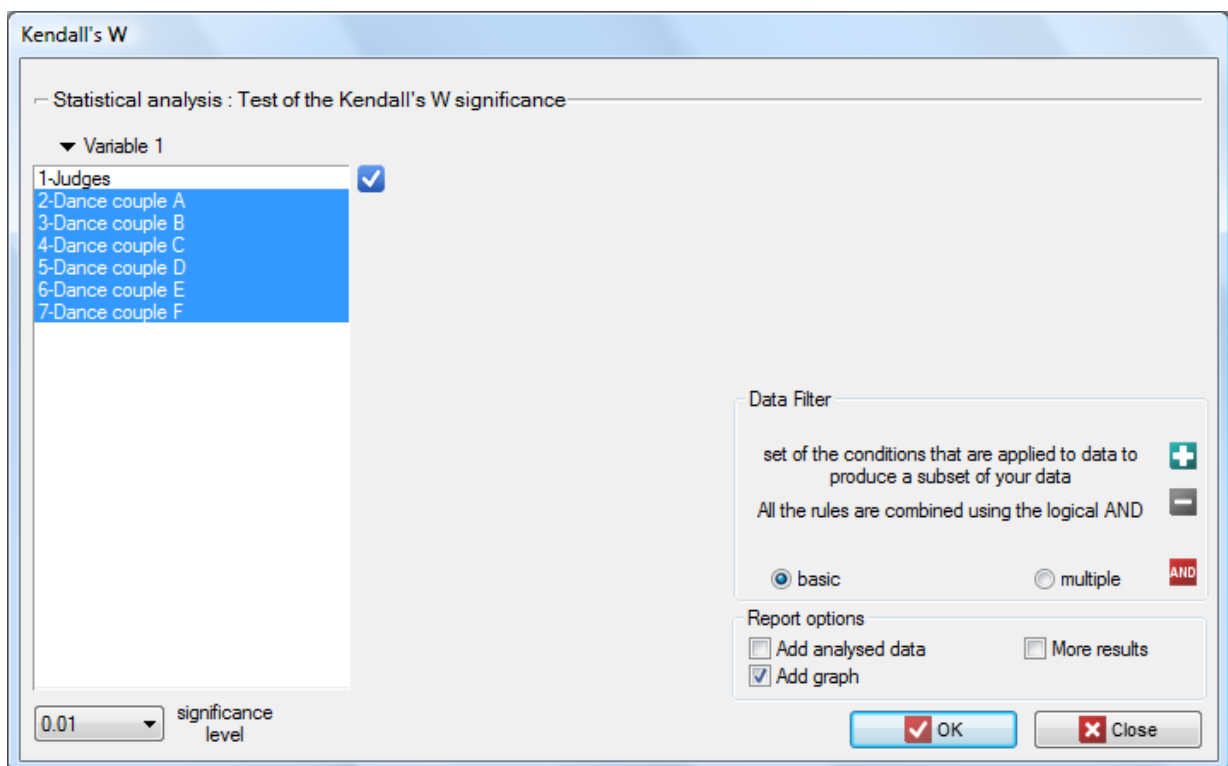
$$\chi^2 = n(k-1)\widetilde{W}$$

This statistic asymptotically (for large sample sizes) has the [rozklad \$\chi^2\$](#) distribution with the degrees of freedom calculated according to the following formula: $df = k - 1$.

The [p value](#), designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

The settings window with the test of the Kendall's W significance can be opened in Statistics menu → NonParametric tests (ordered categories) → Kendall's W or in [Wizard](#).



EXAMPLE 15.2. (judges.pqs file)

In the 6.0 system, dancing pairs grades are assessed by 9 judges. The judges point for example an artistic expression. They assess dancing pairs without comparing each of them and without placing them in the particular "podium place" (they create a ranking). Let's check if the judges assessments are concordant.

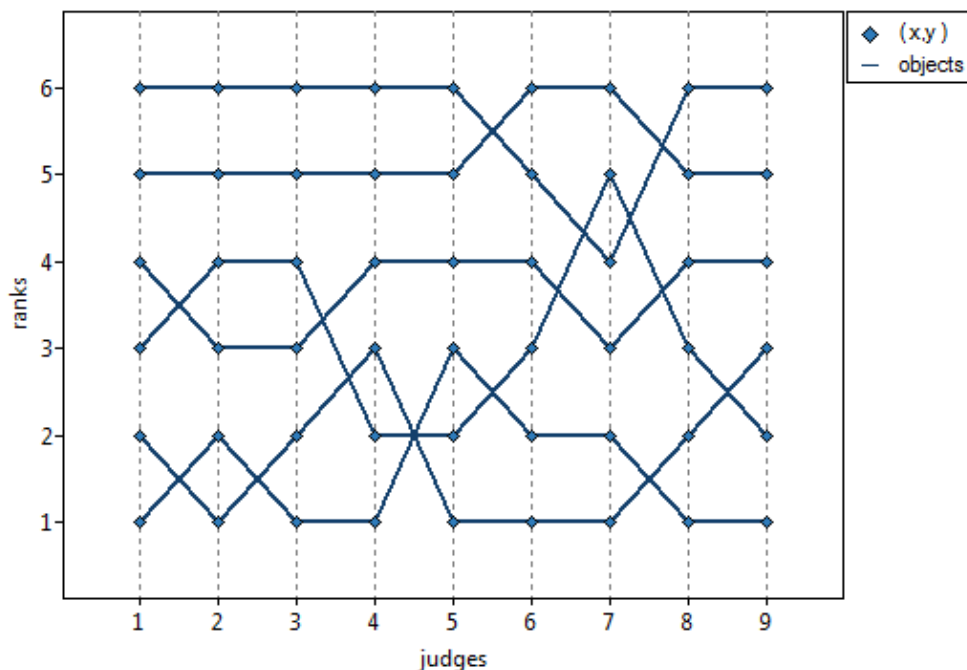
Judges	Couple A	Couple B	Couple C	Couple D	Couple E	Couple F
S1	3	6	2	5	4	1
S2	4	6	1	5	3	2
S3	4	6	2	5	3	1
S4	2	6	3	5	4	1
S5	2	6	1	5	4	3
S6	3	5	1	6	4	2
S7	5	4	1	6	3	2
S8	3	6	2	5	4	1
S9	2	6	3	5	4	1

Hypotheses:

- \mathcal{H}_0 : a lack of concordance between 9 judges assessments,
in the population represented by the sample,
- \mathcal{H}_1 : the 9 judges assessments in the population represented
by the sample are concordant.

Test of the Kendall's W significance	
Analysis time	0.02sec.
Analysed variables	Dance couple A,Dance couple B,Dance couple C,Dance couple D,I
Significance level	0.05
Degrees of freedom	5
Kendall's coefficient of concordance	0.83351
Mean Spearman correlation coefficient	0.812698
Chi2 statistic (adjusted for ties)	37.507937
p-value	<0.000001

Comparing the $p < 0,000001$ with the significance level $\alpha = 0.05$, we have stated that the judges assessments are statistically concordant. The concordance strength is high: $\widetilde{W} = 0.83351$, similarly the average Spearman's rank-order correlation coefficient: $\bar{r}_s = 0.81270$. This result can be presented in the graph, where the X-axis represents the successive judges. Then the more intersection of the lines we can see (the lines should be parallel to the X axis, if the concordance is perfect), the less there is the concordance of judges evaluations.



15.2.2 The Cohen's Kappa coefficient and the test of its significance

The Cohen's Kappa coefficient (Cohen J. (1960)[22]) defines the agreement level of two-times measurements of the same variable in different conditions. Measurement of the same variable can be performed by 2 different observers (reproducibility) or by a one observer twice (recurrence). The $\hat{\kappa}$ coefficient is calculated for categorial dependent variables and its value is included in a range from -1 to 1. A 1 value means a full agreement, 0 value means agreement on the same level which would occur

for data spread in a [contingency table](#) randomly. The level between 0 and -1 is practically not used. The negative $\hat{\kappa}$ value means an agreement on the level which is lower than agreement which occurred for the randomly spread data in a contingency table. The $\hat{\kappa}$ coefficient can be calculated on the basis of [raw data](#) or a $c \times c$ contingency table.

To calculate the $\hat{\kappa}$ coefficient, you need to transform a contingency table for the observed frequencies O_{ij} ([11.6](#)) into the contingency table of probabilities p_{ij} ([15.1](#)):

Table 15.1. The $c \times c$ contingency table of probabilities

Probabilities		$X^{(2)}$				
p_{ij}		$X_1^{(2)}$	$X_2^{(2)}$...	$X_c^{(2)}$	Total
$X^{(1)}$	$X_1^{(1)}$	p_{11}	p_{12}	...	p_{1c}	$p_{1.}$
	$X_2^{(1)}$	p_{21}	p_{22}	...	p_{2c}	$p_{2.}$

	$X_c^{(1)}$	p_{c1}	p_{c2}	...	p_{cc}	$p_{c.}$
	Total	$p_{.1}$	$p_{.2}$...	$p_{.c}$	n

The Kappa coefficient ($\hat{\kappa}$) is defined by:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e},$$

where:

$$P_o = \sum_{i=1}^c p_{ii},$$

$$P_e = \sum_{i=1}^c p_{i.} p_{.i},$$

or equivalently $\hat{\kappa} = (\sum O_{ii} - \sum E_{ii}) / (n - \sum E_{ii})$, where O_{ii} , E_{ii} are the [observed frequencies](#) and the [expected frequencies](#) of main diagonal.

Note

$\hat{\kappa}$ – the coefficient of an agreement in a sample;

κ – the coefficient of an agreement in a population.

The **standard error of Kappa** (Hanley 1987[[38](#)]) is defined by:

$$SE_{\hat{\kappa}} = \frac{\sqrt{A + B - C}}{(1 - P_e)\sqrt{n}},$$

where:

$$A = \sum_{i=1}^c p_{ii} (1 - (p_{i.} + p_{.i})(1 - \hat{\kappa}))^2,$$

$$B = (1 - \hat{\kappa})^2 \sum \sum_{i \neq j} p_{ij} (p_{i.} + p_{.j})^2,$$

$$C = (\hat{\kappa} - P_e(1 - \hat{\kappa}))^2.$$

The **Z test of significance for the Cohen's Kappa** ($\hat{\kappa}$) (Fleiss, 1981[[30](#)]) is used to verify the hypothesis informing us about the agreement of the results of two-times measurements $X^{(1)}$ and $X^{(2)}$ features X and it is based on the $\hat{\kappa}$ coefficient calculated for the sample.

Basic assumptions:

- measurement on a [nominal scale](#) (alternatively: an [ordinal](#) or an [interval](#)).

Hypotheses:

$$\mathcal{H}_0 : \kappa = 0,$$

$$\mathcal{H}_1 : \kappa \neq 0.$$

The test statistic is defined by:

$$Z = \frac{\hat{\kappa}}{SE_{\kappa_{distr}}},$$

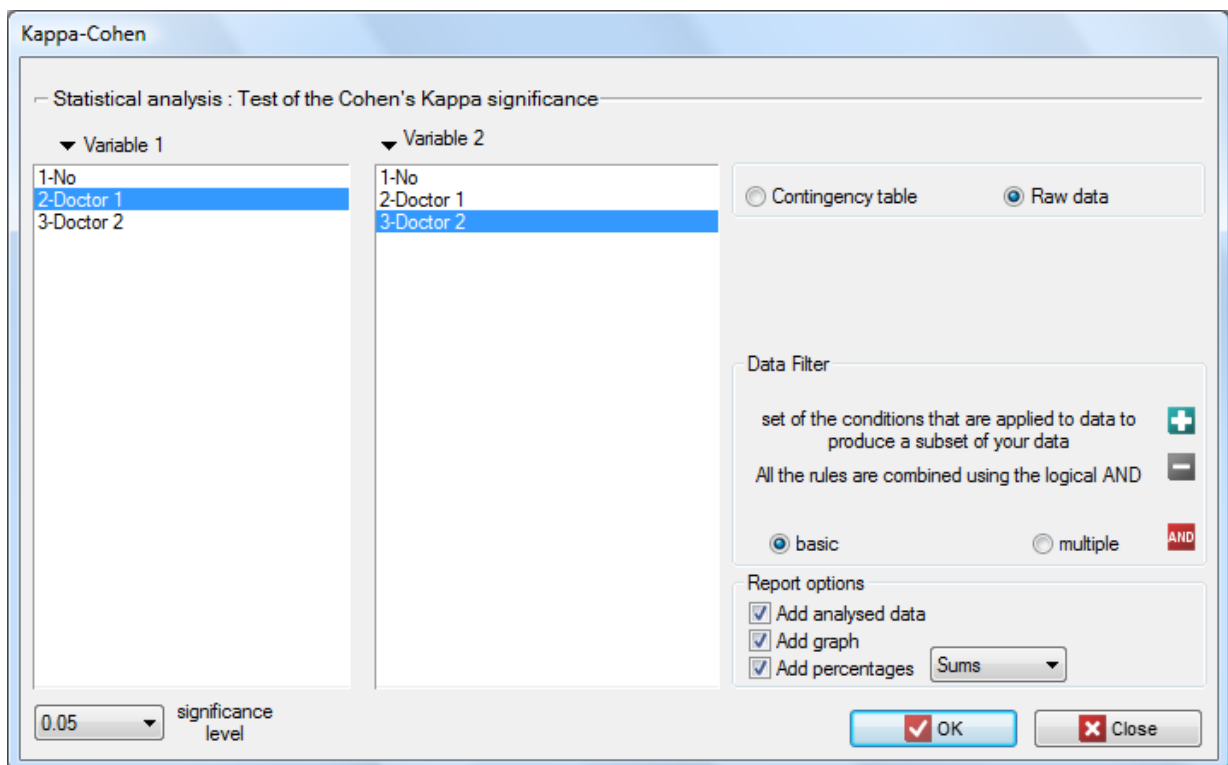
where: $SE_{\kappa_{distr}} = \frac{P_e + P_e^2 - \sum_{i=1}^c p_{i.} p_{.i} (p_{i.} + p_{.i})}{(1 - P_e)^2 n}$ - standard error of a sample distribution.

The Z statistic asymptotically (for a large sample size) has the [normal distribution](#).

The p value, designated on the basis of the [test statistic](#), is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

The settings window with the test of Cohen's Kappa significance can be opened in Statistics menu → NonParametric tests (unordered categories) → Cohen's Kappa or in [Wizard](#).



EXAMPLE 15.3. (diagnosis.pqs file)

You want to analyse the compatibility of a diagnosis made by 2 doctors. To do this, you need to draw 110 patients (children) from a population. The doctors treat patients in a neighbouring doctors' offices. Each patient is examined first by the doctor A and then by the doctor B. Both diagnoses, made by the doctors, are shown in the table below.

	pneumonia	bronchitis	others
pneumonia	31	4	4
bronchitis	8	39	9
others	5	7	3

Hypotheses:

$$\mathcal{H}_0 : \kappa = 0,$$

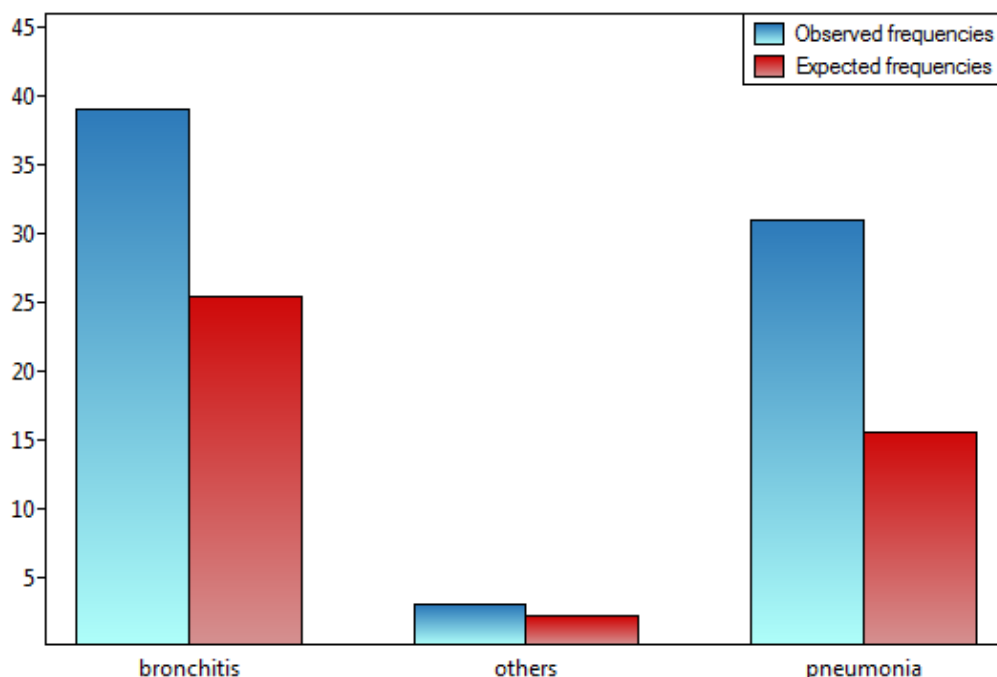
$$\mathcal{H}_1 : \kappa \neq 0.$$

We could analyse the agreement of the diagnoses using just the percentage of the compatible values. In this example, the compatible diagnoses were made for 73 patients (31+39+3=73) which is 66.36% of the analysed group. The kappa coefficient introduces the correction of a chance agreement (it takes into account the agreement occurring by chance).

Test of the Cohen's Kappa significance	
Analysis time	0.04sec.
Analysed variables	Doctor 1;Doctor 2
Significance level	0.05
Size = number of pairs	110
Kappa coefficient	0.4458061
Std. err. of Kappa	0.068029
-95% CI for Kappa coefficient	0.3124717
+95% CI for Kappa coefficient	0.5791405
Std. err. of Kappa distribution	0.0723248
Z statistic	6.163942
p-value (asymptotic)	<0.0000001

Data:	bronchitis	others	pneumonia
bronchitis	39	9	8
others	7	3	5
pneumonia	4	4	31

Sums:	bronchitis	others	pneumonia
bronchitis	35.45%	8.18%	7.27%
others	6.36%	2.73%	4.55%
pneumonia	3.64%	3.64%	28.18%



The agreement with a chance adjustment $\hat{\kappa} = 44,58\%$ is smaller than the one which is not adjusted for the chances of an agreement.

The p value < 0.000001 . Such result proves an agreement between these 2 doctors' opinions, on the significance level $\alpha = 0.05$.

16 DIAGNOSTIC TESTS

16.1 EVALUATION OF DIAGNOSTIC TEST

Suppose that using a diagnostic test we calculate the occurrence of a particular feature (most often disease) and know the gold-standard, so we know that the feature really occurs among the examined people. On the basis of these information, we can build a 2×2 contingency table:

Observed frequencies		Reality (gold-standard)		
		disease (+)	disease free (–)	Total
diagnostic test	positive result (+)	TP	FP	TP+FP
	negative result (–)	FN	TN	FN+TN
	Total	TP+FN	FP+TN	n=TP+FP+FN+TN

where:

TP – true positive

FP – false positive

FN – false negative

TN – true negative

For such a table we can calculate the following measurements.

- **Sensitivity and specificity of diagnostic test**

Every diagnostic test, in some cases, can obtain results different than actual results, for example a diagnostic test, basing on the obtained parameters, classifies a patient to the group of people suffering from a particular disease, or to the group of healthy people. In reality, the number of people approved for the above groups by the test may differ from the number of people genuinely ill and genuinely healthy.

There are two evaluation measurements of the test accuracy. They are:

Sensitivity – describes the ability to detect people genuinely ill (having a particular feature). If we examine a group of ill people, the sensitivity provides us with the information what percentage of them have a positive test result.

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

Confidence interval is built on the basis of the [Clopper-Pearson](#) method for a single proportion.

Specificity – describes the ability to detect people genuinely healthy (without a particular feature). If we examine a group of genuinely healthy people, the specificity provides us with the information about the percentage of people having a negative test result.

$$\text{specificity} = \frac{TN}{FP + TN}$$

Confidence interval is built on the basis of the [Clopper-Pearson](#) method for a single proportion.

- **Positive predictive values, negative predictive values and prevalence rate**

Positive predictive value (PPV) – the probability, that a person having a positive test result suffered from a disease. If the examined person obtains a positive test result, the PPV informs them how they can be sure, that they suffer from a particular disease.

$$PPV = \frac{TP}{TP + FP}$$

Confidence interval is built on the basis of the [Clopper-Pearson](#) method for a single proportion.

Negative predictive value (NPV) – the probability that a person having a negative test result did not suffer from any disease. If the examined person obtains a negative test result, the NPV informs them how they can be sure that they do not suffer from a particular disease.

$$NPV = \frac{TN}{FN + TN}$$

Confidence interval is built on the basis of the [Clopper-Pearson](#) method for a single proportion.

Positive and negative predictive values depend on the prevalence rate.

Prevalence – probability of disease in the population for which the diagnostic test was conducted.

$$\text{prevalence} = \frac{TP + FN}{n}$$

Confidence interval is built on the basis of the [Clopper-Pearson](#) method for a single proportion.

- **Likelihood ratio of positive test and likelihood ratio of negative test**

Likelihood ratio of positive test (LR_+) – this measurement enables the comparison of some test results matching to the gold-standard. It does not depend on the prevalence of the disease. It is the ratio of two odds: the odds that a person from the group of ill people will obtain a positive test result, and the same effect will be observed among healthy people.

$$LR_+ = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{TP(TP + FN)}{FP(FP + TN)}$$

Confidence interval for LR_+ is built on the basis of the standard error:

$$SE = \sqrt{\frac{1 - \text{sensitivity}}{TP} + \frac{\text{specificity}}{FP}}.$$

Likelihood ratio of negative test (LR_-) – it is the ratio of two odds: the odds that a person from the group of ill people will obtain a negative test result, and the same effect will be observed among healthy people.

$$LR_- = \frac{1 - \text{sensitivity}}{\text{specificity}} = \frac{FN(TP + FN)}{TN(FP + TN)}$$

Confidence interval for LR_- is built on the basis of the standard error:

$$SE = \sqrt{\frac{\text{sensitivity}}{FN} + \frac{1 - \text{specificity}}{TN}}.$$

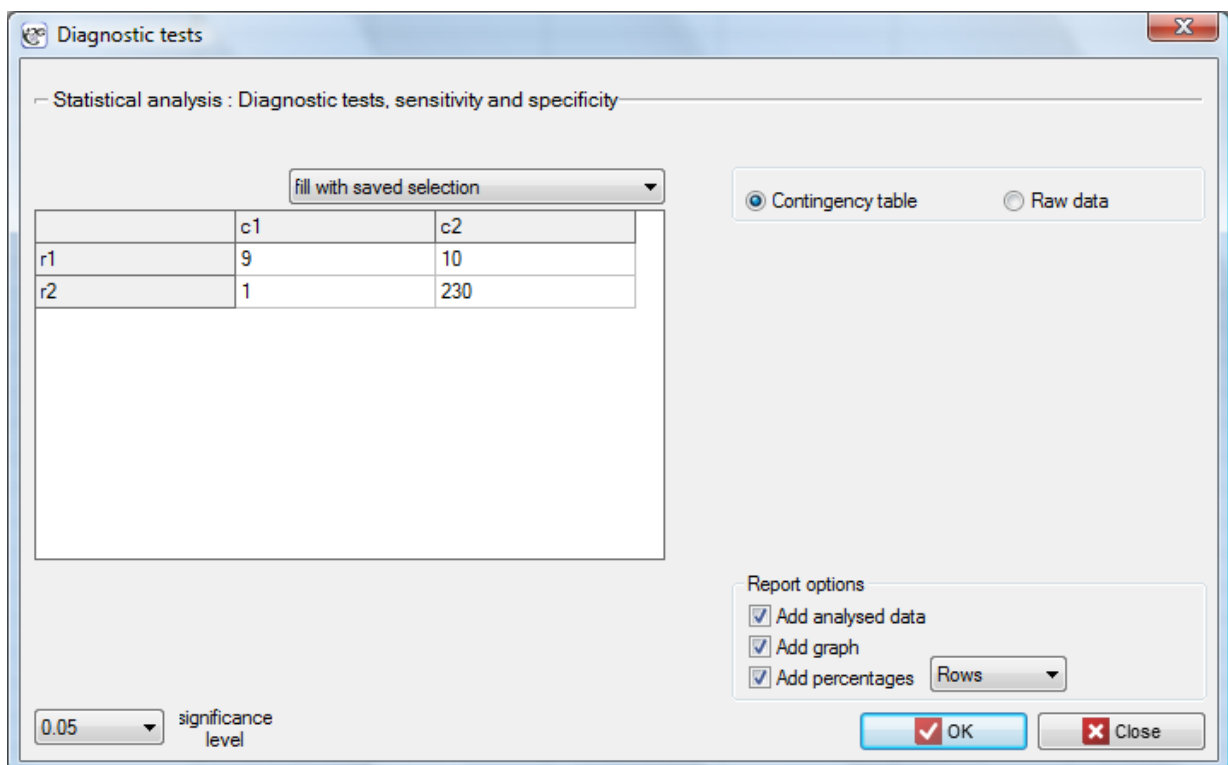
- **Accuracy**

Accuracy (Acc) – the probability of a correct diagnose using a diagnostic test. If the examined person obtains a positive or a negative test result, the Acc informs how they can be sure about the definitive diagnosis.

$$Acc = \frac{TP + TN}{n}$$

Confidence interval is built on the basis of the **Clopper-Pearson** method for a single proportion.

The settings window with the diagnostic tests can be opened in Statistics menu → Diagnostic tests → Diagnostic tests



EXAMPLE 16.1. (mammography.pqs file)

Mammography is one of the most popular screening tests which enables the detection of breast cancer. The following study has been carried out on the group of 250 people, so-called "asymptomatic" women at the age from 40 to 50. Mammography can detect an outbreak of cancer smaller than 5 mm and enables to note the change which is not a nodule yet but a change in the structure of tissues.

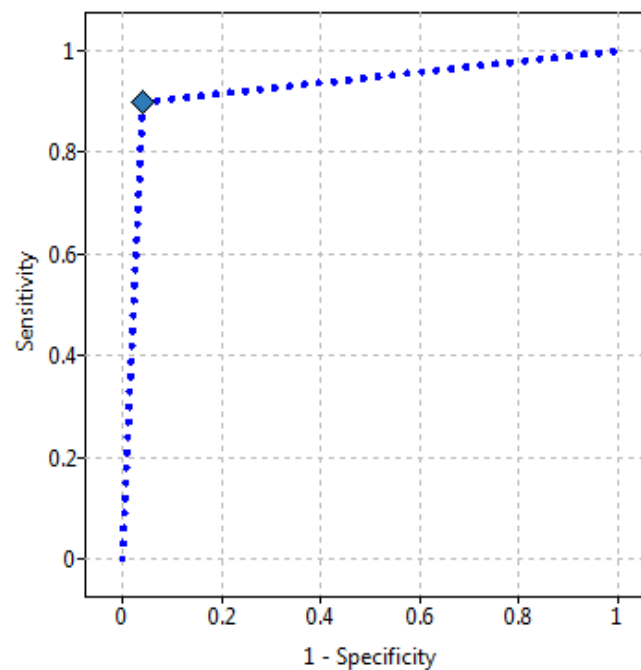
Observed frequencies		Reality (histopatology)		
		disease (+)	disease free (–)	Total
mammography	positive result (+)	9	10	19
	negative result (–)	1	230	231
	Total	10	240	250

We will calculate the values enabling the assessment of the performed diagnostic test.

Diagnostic tests, sensitivity and specificity	
Analysis time	0.50sec.
Analysed variables	Contingency table
Significance level	0.05
Sensitivity	0.9
-95% CI	0.554984
+95% CI	0.997471
Specificity	0.958333
-95% CI	0.92471
+95% CI	0.979841
Positive predictive value (PPV)	0.473684
-95% CI	0.244475
+95% CI	0.711357
Negative predictive value (NPV)	0.995671
-95% CI	0.976118
+95% CI	0.99989
Positive likelihood ratio (PLR)	21.6
-95% CI	11.37865
+95% CI	41.003106
Negative likelihood ratio (NLR)	0.104348
-95% CI	0.016251
+95% CI	0.670016
Accuracy (ACC)	0.956
-95% CI	0.922637
+95% CI	0.977834
Prevalence	0.04
-95% CI	0.019345
+95% CI	0.072329

Data:		
	c1	c2
r1	9	10
r2	1	230

Rows:		
	c1	c2
r1	47.37%	52.63%
r2	0.43%	99.57%



- 90% of women suffering from breast cancer have been correctly defined, so they have obtained the positive result of mammography;

- 95.83% of healthy women (not suffering from breast cancer) have been correctly defined, so they have obtained the negative result of mammography;
- 4 out of 100 examined women suffer from breast cancer;
- A woman who have obtained a positive mammography result can be 47.37% sure that she suffers from breast cancer;
- A women who have obtained a negative test result can be 99.57% sure that she does not suffer from breast cancer;
- The probability that the positive mammography result will be obtained by a woman genuinely suffering from cancer is 21.60 times greater than the probability that the positive mammography result will be obtained by a healthy woman (not suffering from breast cancer);
- The probability that the negative mammography result will be obtained by a woman genuinely suffering from breast cancer is 10.43% of the probability that the negative mammography result will be obtained by a healthy woman (not suffering from breast cancer);
- A woman undergoing mammography (regardless of age) can be 96.50% sure of the definitive diagnosis.

16.2 ROC CURVE

The diagnostic test is used for differentiating objects with a given feature (marked as **(+)**, e.g. ill people) from objects without the feature (marked as **(-)**, e.g. healthy people). For the diagnostic test to be considered valuable, it should yield a relatively small number of wrong classifications. If the test is based on a dichotomous variable then the proper tool for the evaluation of the quality of the test is the analysis of a 2×2 contingency table of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Most frequently, though, diagnostic tests are based on continuous variables or ordered categorical variables. In such a situation the proper means of evaluating the capability of the test for differentiating **(+)** and **(-)** are ROC (Receiver Operating Characteristic) curves.

It is frequently observed that the greater the value of the diagnostic variable, the greater the odds of occurrence of the studied phenomenon, or the other way round: the smaller the value of the diagnostic variable, the smaller the odds of occurrence of the studied phenomenon. Then, with the use of ROC curves, the choice of the optimum cut-off is made, i.e. the choice of a certain value of the diagnostic variable which best separates the studied statistical population into two groups: **(+)** in which the given phenomenon occurs and **(-)** in which the given phenomenon does not occur.

When, on the basis of the studies of the same objects, two or more ROC curves are constructed, one can compare the curves with regard to the quality of classification.

Let us assume that we have at our disposal a sample of n elements, in which each object has one of the k values of the diagnostic variable. Each of the received values of the diagnostic variable x_1, x_2, \dots, x_k becomes the cut-off x_{cat} .

If the diagnostic variable is:

- **stimulant** (the growth of its value makes the odds of occurrence of the studied phenomenon greater), then values greater than or equal to the cut-off ($x_i \geq x_{cat}$) are classified in group **(+)**;
- **destimulant** (the growth of its value makes the odds of occurrence of the studied phenomenon smaller), then values smaller than or equal to the cut-off ($x_i \geq x_{cat}$) are classified in group **(+)**;

For each of the k cut-offs we define true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

stimulant		Reality	
		(+)	(-)
diagnostic variable	$x_i \geq x_{cat}$ (+)	TP	FP
	$x_i < x_{cat}$ (-)	FN	TN

destimulant		Reality	
		(+)	(-)
diagnostic variable	$x_i \leq x_{cat}$ (+)	TP	FP
	$x_i > x_{cat}$ (-)	FN	TN

On the basis of those values each cut-off x_{cat} can be further described by means of [sensitivity](#) and [specificity](#), positive predictive values ([PPV](#)), negative predictive values ([NPV](#)), positive result likelihood ratio ([LR₊](#)), negative result likelihood ratio ([LR₋](#)), and accuracy ([Acc](#)).

Note

The PQStat program computes the prevalence coefficient on the basis of the sample. The computed prevalence coefficient will reflect the occurrence of the studied phenomenon (illness) in the population in the case of screening of a large sample representing the population. If only people with suspected illness are directed to medical examinations, then the computed prevalence coefficient for them can be much higher than the prevalence coefficient for the population.

Because both the positive and negative predictive value depend on the prevalence coefficient, when the coefficient for the population is known a priori, we can use it to compute, for each cut-off x_{cat} , corrected predictive values according to Bayes's formulas:

$$PPV_{revised} = \frac{\text{Sensitivity} \cdot P_{apriori}}{\text{Sensitivity} \cdot P_{apriori} + (1 - \text{Specificity}) \cdot (1 - P_{apriori})}$$

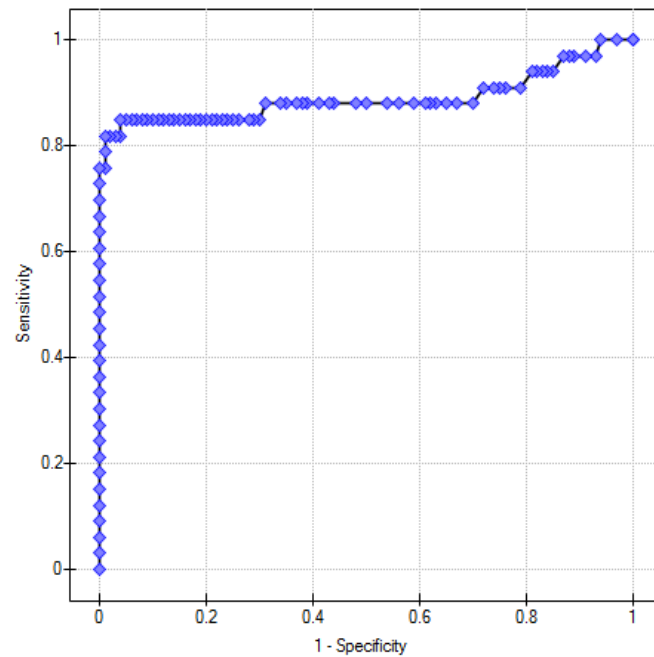
$$NPV_{revised} = \frac{\text{Specificity} \cdot (1 - P_{apriori})}{\text{Specificity} \cdot (1 - P_{apriori}) + (1 - \text{Sensitivity}) \cdot P_{apriori}}$$

where:

$P_{apriori}$ - the prevalence coefficient put in by the user, the so-called pre-test probability of disease

x_{cat}	sensitivity	specificity	PPV	NPV	LR ₊	LR ₋	Acc	PPV _{rev}	NPV _{rev}
x_1	sensitivity ₁	specificity ₁	PPV ₁	NPV ₁	LR ₊₁	LR ₋₁	Acc ₁	PPV _{rev1}	NPV _{rev1}
x_2	sensitivity ₂	specificity ₂	PPV ₂	NPV ₂	LR ₊₂	LR ₋₂	Acc ₂	PPV _{rev2}	NPV _{rev2}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_k	sensitivity _k	specificity _k	PPV _k	NPV _k	LR _{+k}	LR _{-k}	Acc _k	PPV _{revk}	NPV _{revk}

The ROC curve is created on the basis of the calculated values of sensitivity and specificity. On the abscissa axis the $x=1-\text{specificity}$ is placed, and on the ordinate axis $y=\text{sensitivity}$. The points obtained in that manner are linked. The constructed curve, especially the area under the curve, presents the classification quality of the analyzed diagnostic variable. When the ROC curve coincides with the diagonal $y = x$, then the decision made on the basis of the diagnostic variable is as good as the random distribution of studied objects into group (+) and group (-).



AUC (area under curve) – the size of the area under the ROC curve falls within $< 0; 1 >$. The greater the field the more exact the classification of the objects in group (+) and group (–) on the basis of the analyzed diagnostic variable. Therefore, that diagnostic variable can be even more useful as a classifier. The area AUC , error SE_{AUC} and confidence interval for AUC are calculated on the basis of:

- ★ nonparametric **DeLong** method (DeLong E.R. et al. 1988[26], Hanley J.A. i Hajian-Tilaki K.O. 1997[38]) - **recommended**,
- ★ nonparametric **Hanley-McNeil** method (Hanley J.A. i McNeil M.D. 1982[39]),
- ★ **Hanley-McNeil** method which presumes double negative exponential distribution (Hanley J.A. i McNeil M.D. 1982[39]) - computed only when groups (+) and (–) are equinumerous.

For the classification to be better than random distribution of objects into to classes, the area under the ROC curve should be significantly larger than the area under the line $y = x$, i.e. than 0.5.

Hypotheses:

$$\mathcal{H}_0 : AUC = 0.5,$$

$$\mathcal{H}_1 : AUC \neq 0.5.$$

The test statistics has the form presented below:

$$Z = \frac{AUC - 0.5}{SE_{0.5}},$$

where:

$$SE_{0.5} = \sqrt{\frac{n_{(+)} + n_{(-)} + 1}{12n_{(+)}n_{(-)}}},$$

$n_{(+)}$ – size of the sample (+) in which the given phenomenon occurs,

$n_{(-)}$ – size of the sample (–), in which the given phenomenon does not occur.

The Z statistic asymptotically (for large sample sizes) has the **normal distribution**.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

if $p \leq \alpha \implies$ reject \mathcal{H}_0 and accept \mathcal{H}_1 ,

if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

16.2.1 Selection of optimum cut-off

The point which is looked for is a certain value of the diagnostic variable, which provides the optimum separation of the studied population into two groups: **(+)** in which the given phenomenon occurs and **(-)** in which the given phenomenon does not occur. The selection of the optimum cut-off is not easy because it requires specialist knowledge about the topic of the study. For example, different cut-offs will be required in, on the one hand, a test used for screening of a large group of people, e.g. for a mammography study, and, on the other hand, in invasive studies conducted for the purpose of confirming an earlier suspicion, e.g. in histopathology. With the help of an advanced mathematical apparatus we can find a cut-off which will be the most useful from the perspective of mathematics.

PQStat Program enables the selection of an optimum cut-off by means of an analysis of the graph of the intersection of sensitivity and specificity. Besides, the optimum cut-off can be computed on the basis of data about the costs of wrong decisions and about the a priori prevalence coefficient value, provided by the user.

- **Optimum cut-off on ROC curve** — computed on the basis of sensitivity, specificity, costs of wrong decisions, and the prevalence coefficient.

Errors which can be made when classifying the studied objects as belonging to group **(+)** and group **(-)** are false positive results (*FP*) and false negative results (*FN*). If committing those errors is equally costly (ethical, financial, and other costs), then in the field Cost *FP* and in the field Cost *FN* we enter the same positive value — usually 1. However, if we come to the conclusion that one type of error is encumbered with a greater cost than the other one, then we will assign appropriately greater weight to it.

The optimum cut-off value is calculated on the basis of sensitivity, specificity, and with the help of value m — slope of the tangent line to the ROC curve. The slope angle m is defined in relation to two values: the costs of wrong decisions and the prevalence coefficient. Normally the costs of wrong decisions have the value 1 and the prevalence coefficient is estimated from the sample. Knowing, a priori, the prevalence coefficient ($P_{\text{a priori}}$) and the costs of wrong decisions, the user can influence the value m and, consequently, the search for an optimum cut-off. As a result, the optimum cut-off is determined to be such a value of the diagnostic variable for which the formula:

$$\text{Sensitivity} - m \cdot (1 - \text{Specificity})$$

reaches the minimum (Zweig M.H. 1993[89]).

The optimum cut-off point of the diagnostic variable, selected as described above, will finally be marked on the ROC curve.

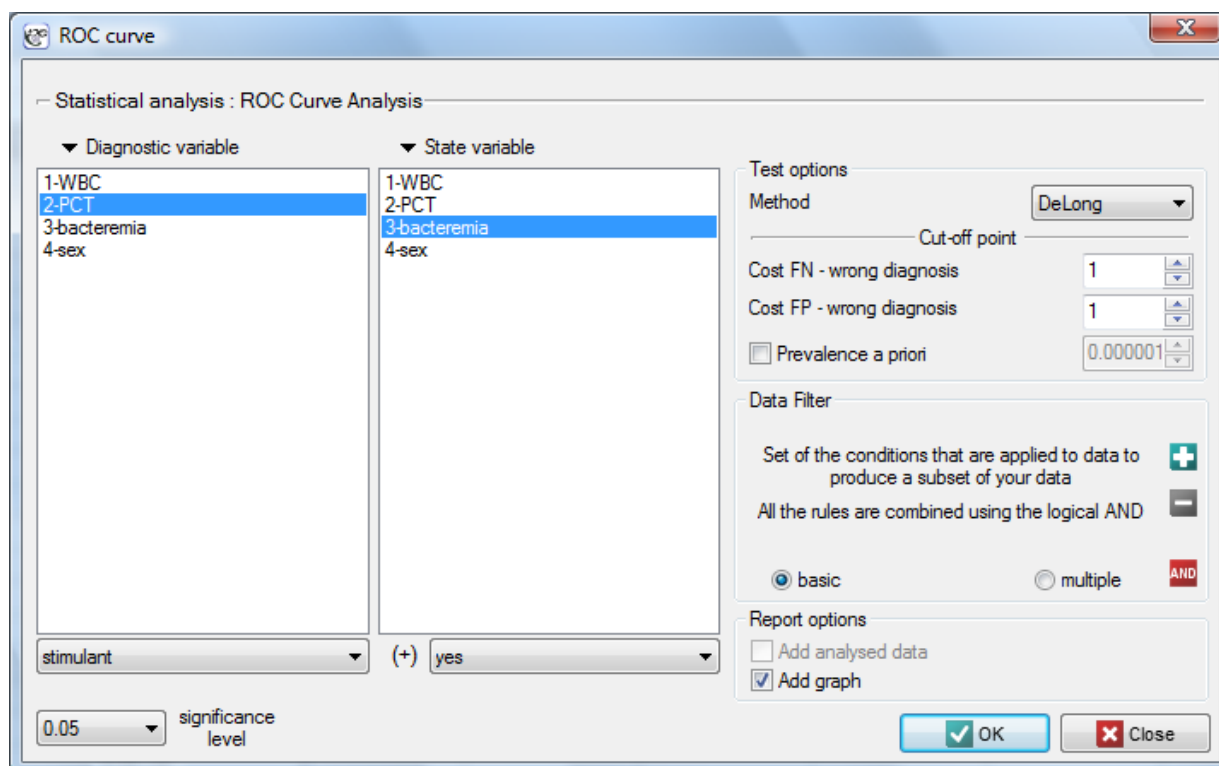
- **Costs graph** — presents the calculated values of an wrong diagnosis together with their costs. The values are computed according to the formula:

$$\text{cost} = \text{cost}_{FP} \cdot FP + \text{cost}_{FN} \cdot FN$$

The point marked on the graph is the minimum of the function presented above.

- **Sensitivity and specificity intersection graph** — allows the localization of the point in which the value of sensitivity and specificity is simultaneously the greatest.

The window with settings for ROC analysis is accessed via the menu Statistics → Diagnostic tests → ROC curve.

**EXAMPLE 16.2.** (file bacteriemia.pqs)

Persistent high fever in an infant or a small child without clearly diagnosed reasons is a premise for testing for bacteremia. The most useful and reliable parameters for screening and monitoring bacterial infections are the following indicators:

WBC - the number of white blood cells

PCT - procalcitonin.

It is assumed that in a healthy infant or a small child WBC should not exceed 15 thousand/ μ l and PCT should be lower than 0.5 ng/ml.

The sample values of those indicators for 136 children of up to 3 years old with persistent fever $> 39^{\circ}\text{C}$ is presented in the table fragment below:

WBC	PCT	bacteremia	sex
11,30	0,023	no	f
11,00	0,022	no	f
6,70	0,009	no	f
5,90	0,004	no	f
6,10	0,006	no	f
12,50	0,031	no	f
4,90	0,002	no	f
6,90	0,011	no	f
11,60	0,025	no	f
20,90	5,919	yes	f
20,80	6,405	yes	f
0,00	0,017	no	f

One method of analyzing the PCT indicator is transforming it into a dichotomous variable by selecting a cut-off (e.g. $x_{cat}=0.5$ ng/ml) above which the study is considered to be "positive". The level of adequacy of such a division will be indicated by the value of sensitivity and specificity. We want to use a more complex approach, that is, calculate the sensitivity and specificity not only for one value but for each PCT value obtained in the sample - which means constructing a ROC curve. On the basis of the

information obtained in that manner we want to check if the PTC indicator is indeed useful for diagnosing bacteremia. If so, then we want to check what is the optimal cut-off above which we can consider the study to be "positive" – detecting bacteremia.

In order to check if PTC is really useful for diagnosing bacteremia we will calculate the size of the area under the ROC curve and verify the hypothesis that:

$$\mathcal{H}_0 : \text{area under the constructed ROC curve} = 0.5,$$


$$\mathcal{H}_1 : \text{area under the constructed ROC curve} \neq 0.5.$$

As bacteremia is accompanied by an increased PCT level, in the test options window we will consider the indicator to be a stimulant. In the state variable we have to define which value in the bacteremia column determines its presence, then we select "yes". Apart from the result of the statistical test, in the report we can find an exact description of every possible cut-off.

ROC Curve Analysis	
Analysis time	0,29sec.
Analysed variables	PCT;bacteremia
Count of missing data	3
Significance level	0,05
Size	133
Size STATE + (yes)	33
Size STATE - (no)	100
Direction of diagnostic variable	stimulant
Prevalence	0,248120301
-95% CI	0,177358632
+95% CI	0,330437258
DeLong's method	
AUC	0,889242424
SE(AUC)	0,048124092
-95% CI	0,794920921
+95% CI	0,9835639
Z statistic	6,691374893
p-value	<0.000000001

PCT	STATE -	STATE +	FP	TP	TN	FN	Sensitivity	Specificity	PPV	NPV	LR+	LR-	ACC
0,001	3	0	100	33	0	0	1	0	0,2481203	NA	1	NA	0,2481203
0,002	3	0	97	33	3	0	1	0,03	0,2538461	1	1,0309278	0	0,2706766
0,003	3	1	94	33	6	0	1	0,06	0,2598425	1	1,0638297	0	0,2932330
0,004	2	0	91	32	9	1	0,9696969	0,09	0,2601626	0,9	1,0656010	0,3367003	0,3082706
0,005	1	0	89	32	11	1	0,9696969	0,11	0,2644628	0,9166666	1,0895471	0,2754820	0,3233082
0,006	1	0	88	32	12	1	0,9696969	0,12	0,2666666	0,9230769	1,1019283	0,2525252	0,3308270
0,007	3	1	87	32	13	1	0,9696969	0,13	0,2689075	0,9285714	1,1145942	0,2331002	0,3383458
0,008	1	0	84	31	16	2	0,9393939	0,16	0,2695652	0,8888888	1,1183261	0,3787878	0,3533834
0,009	2	0	83	31	17	2	0,9393939	0,17	0,2719298	0,8947368	1,1317999	0,3565062	0,3609022
0,011	2	1	81	31	19	2	0,9393939	0,19	0,2767857	0,9047619	1,1597456	0,3189792	0,3759398
0,012	3	0	79	30	21	3	0,9090909	0,21	0,2752293	0,875	1,1507479	0,4329004	0,3834586
0,013	2	0	76	30	24	3	0,9090909	0,24	0,2830188	0,8888888	1,1961722	0,3787878	0,4060150
0,014	2	0	74	30	26	3	0,9090909	0,26	0,2884615	0,8965517	1,2285012	0,3496503	0,4210526
0,015	2	1	72	30	28	3	0,9090909	0,28	0,2941176	0,9032258	1,2626262	0,3246753	0,4360902
0,016	3	0	70	29	30	4	0,8787878	0,3	0,2929292	0,8823529	1,2554112	0,4040404	0,4436090
0,017	2	0	67	29	33	4	0,8787878	0,33	0,3020833	0,8918918	1,3116236	0,3673094	0,4661654
0,018	2	0	65	29	35	4	0,8787878	0,35	0,3085106	0,8974358	1,3519813	0,3463203	0,4812030

The calculated size of the area under the ROC curve is $AUC = 0.889$. Therefore, on the basis of the adopted level $\alpha = 0.05$, based on the obtained value $p < 0.000001$ we assume that diagnosing bac-

teremia with the use of the PCT indicator is indeed more useful than a random distribution of patients into 2 groups: suffering from bacteremia and not suffering from it. Therefore, we return to the analysis (button ) to define the optimal cut-off.

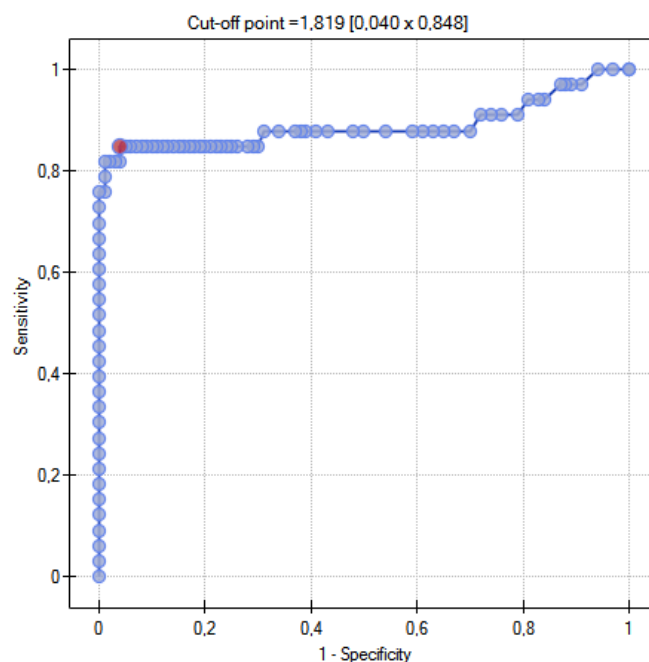
The algorithm of searching for the optimal cut-off takes into account the costs of wrong decisions and the prevalence coefficient.

- (1) FN cost - wrong diagnosis is the cost of assuming that the patient does not suffer from bacteremia although in reality he or she is suffering from it (costs of a falsely negative decision)
- (2) FP cost - wrong diagnosis, is the cost of assuming that the patient suffers from bacteremia although in reality he or she is not suffering from it (costs of a falsely positive decision)

As the FN costs are much more serious than the FP costs, we enter a greater value in field one than in field two. We decided the value would be 5.

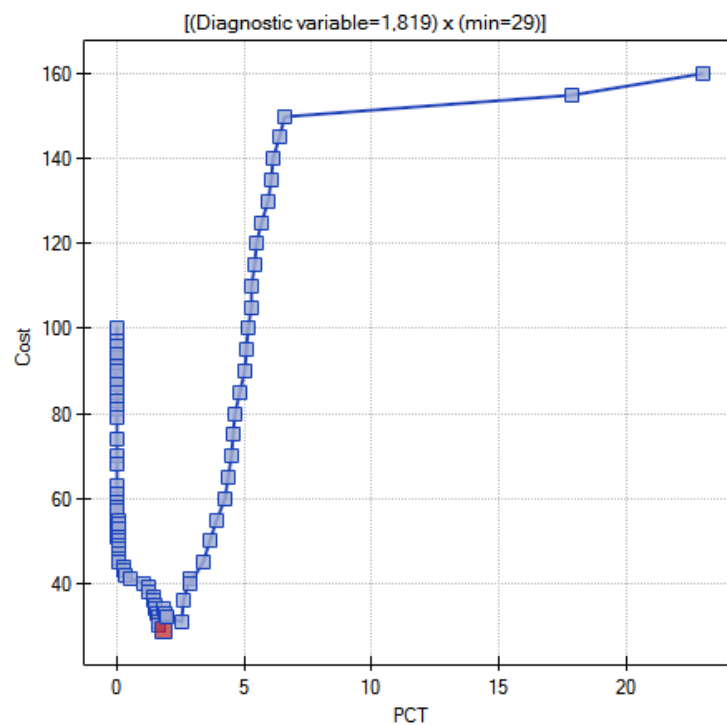
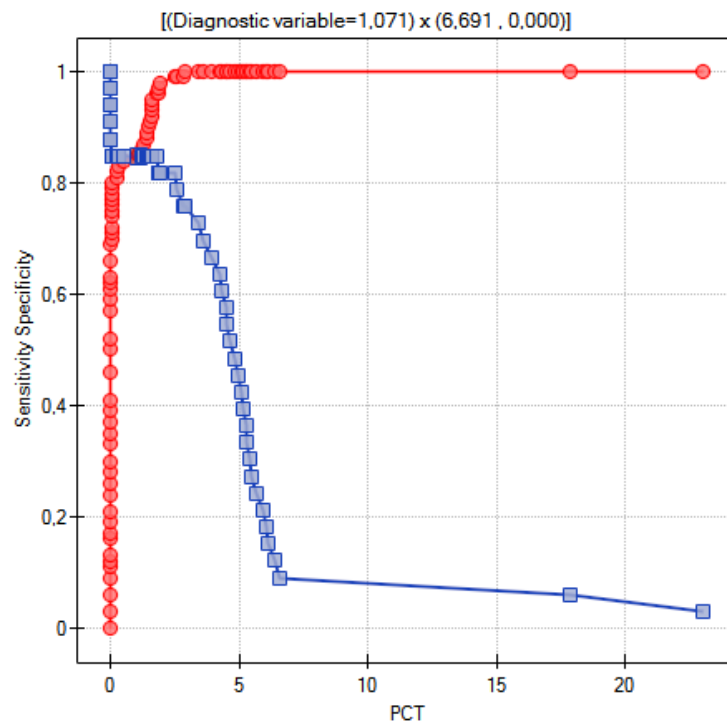
The PCT value is to be used in screening so we do not give the prevalence coefficient for the population (a priori prevalence coefficient) which is very low but we use the estimated coefficient from the sample. We do so in order not to move the cut-off of the PCT value too high and not to increase the number of falsely negative results.

For cut-off	
Cost FN - wrong diagnosis	5
Cost FP - wrong diagnosis	1



The optimal PCT cut-off determined in this way is 1.819. For this point sensitivity=0.85 and specificity=0.96.

Another method of selecting the cut-off is the analysis of the costs graph and of the sensitivity intersection graph:



The analysis of the costs graph shows that the minimum of the costs of wrong decisions lies at $PCT=1.819$. The value of sensitivity and specificity is similar at $PCT=1.071$.

16.2.2 ROC curves comparison

Very often the aim of studies is the comparison of the size of the area under the ROC curve (AUC_1) with the area under another ROC curve (AUC_2). The ROC curve with a greater area usually allows a more precise classification of objects.

Methods for comparing the areas depend on the model of the study.

- **Dependent model** — the compared ROC curves are constructed on the basis of measurements made on the same objects.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : AUC_1 &= AUC_2, \\ \mathcal{H}_1 : AUC_1 &\neq AUC_2.\end{aligned}$$

The test statistics has the form presented below:

$$Z = \frac{|AUC_1 - AUC_2|}{SE_{AUC_1 - AUC_2}},$$

where:

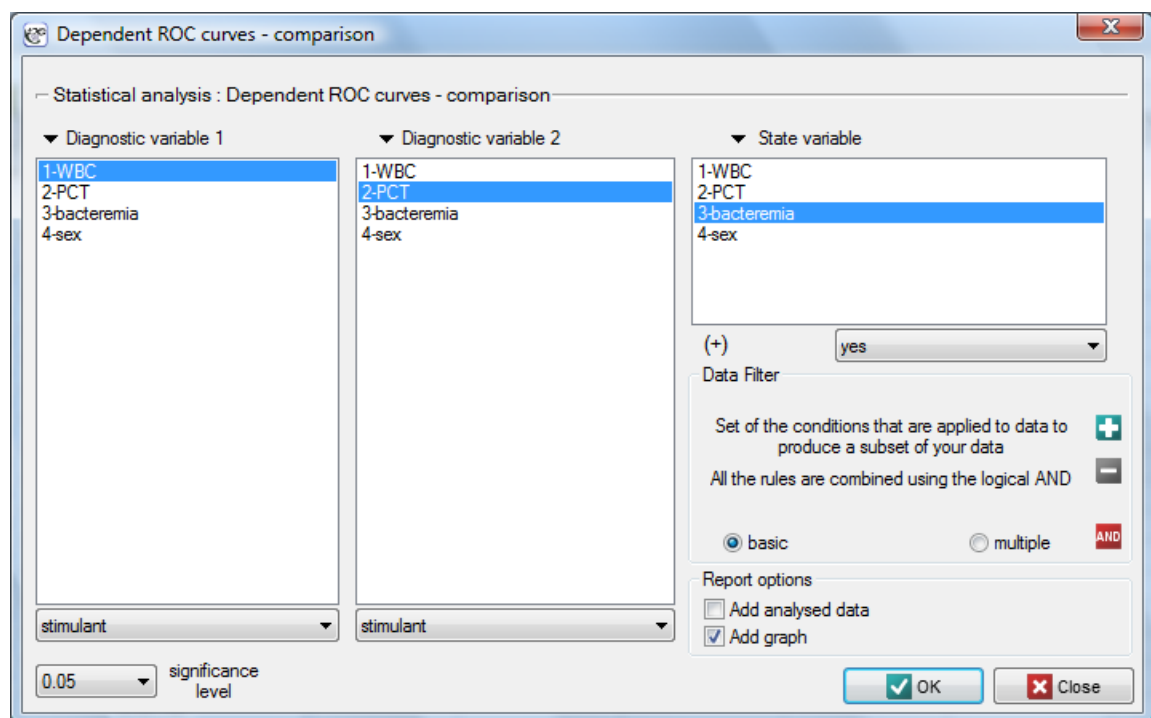
AUC_1 , AUC_2 and the standard error of the difference in areas $SE_{AUC_1 - AUC_2}$ are calculated on the basis of the nonparametric method proposed by **DeLong** (DeLong E.R. et al., 1988[26], Hanley J.A., and Hajian-Tilaki K.O. 1997[38])

Statistics Z has (for large sizes) asymptotic **normal distribution**.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The window with settings for comparing dependent ROC curves is accessed via the menu Statistics→Diagnostic tests→Dependent ROC Curves – comparison.



- **Independent model** — the compared ROC curves are constructed on the basis of measurements made on different objects.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : AUC_1 &= AUC_2, \\ \mathcal{H}_1 : AUC_1 &\neq AUC_2.\end{aligned}$$

Test statistics (Hanley J.A. and McNeil M.D. 1983[40]) has the form:

$$Z = \frac{|AUC_1 - AUC_2|}{\sqrt{SE_{AUC_1}^2 - SE_{AUC_2}^2}},$$

where:

AUC_1 , AUC_2 and standard errors of areas SE_{AUC_1} , SE_{AUC_2} are calculated on the basis of:

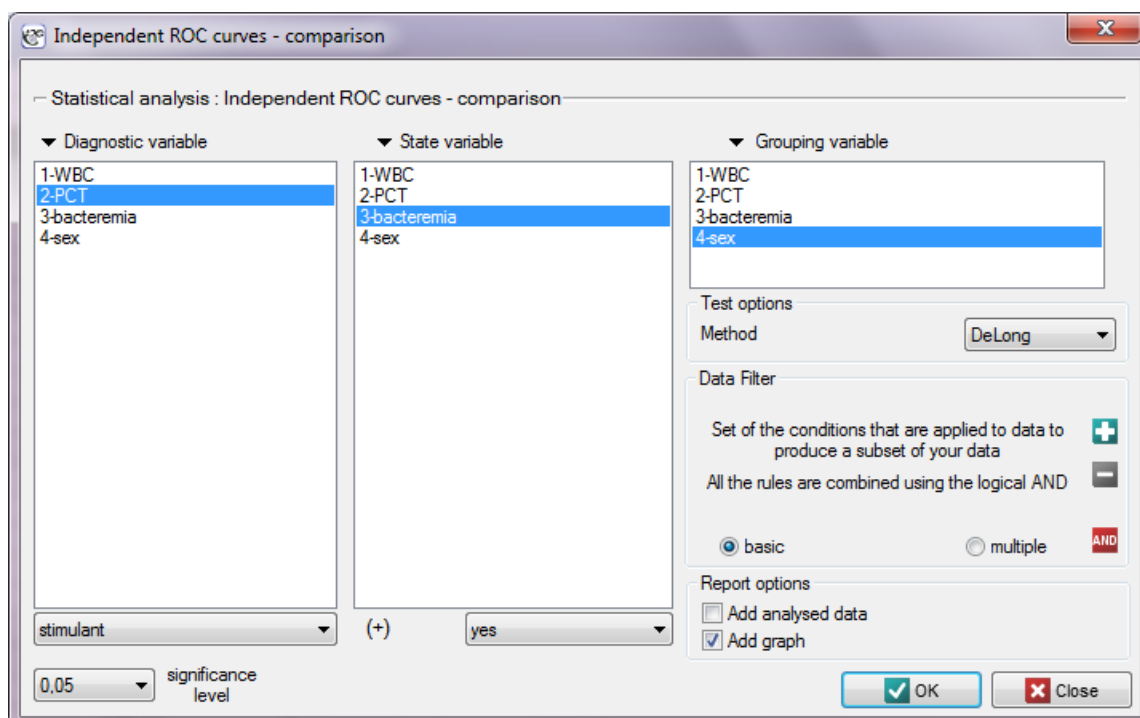
- ★ nonparametric method **DeLong** (DeLong E.R. et al. 1988[26], Hanley J.A., and Hajian-Tilaki K.O. 1997[38]) - **recommended**,
- ★ nonparametric **Hanley-McNeil** method (Hanley J.A. and McNeil M.D. 1982[39]),
- ★ method which presumes double negative exponential distribution (Hanley J.A. and McNeil M.D. 1982[39]) - computed only when groups **(+)** and **(-)** are equinumerous.

Statistics Z has (for large sizes) asymptotic **normal distribution**.

On the basis of **test statistics p value** is estimated and then compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no basis for rejecting } \mathcal{H}_0.\end{aligned}$$

The window with settings for comparing independent ROC curves is accessed via the menu Statistics→Diagnostic tests→Independent ROC Curves – comparison.



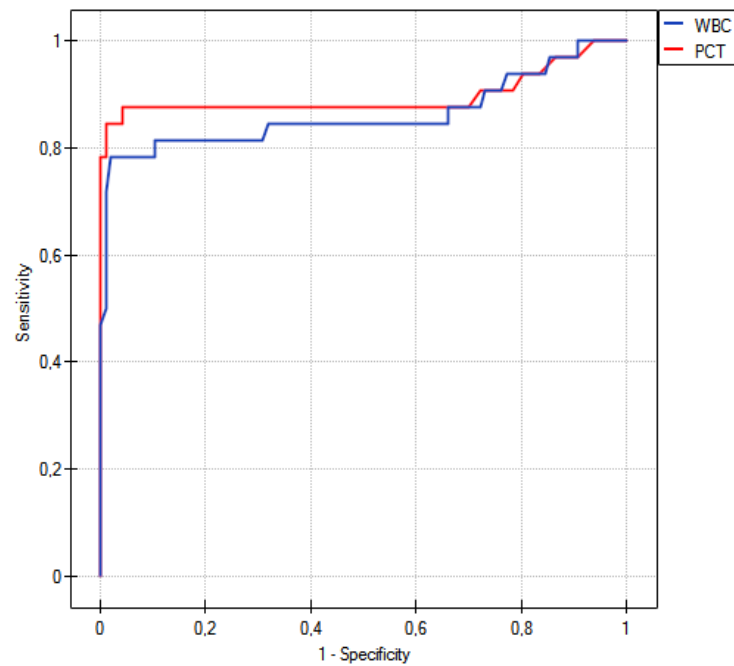
EXAMPLE (16.2) c.d. (bacteriemia.pqs file)

We will make 2 comparisons:

- 1) We will construct 2 ROC curves to compare the diagnostic value of parameters WBC and PCT;
 - 2) We will construct 2 ROC curves to compare the diagnostic value of PCT parameter for boys and girls.
- ad1) Both parameters, WBC and PCT, are stimulants (in bacteremia their values are high). In the course of the comparison of the diagnostic value of those parameters we verify the following hypotheses:

\mathcal{H}_0 : the area under ROC curve for WBC = the area under the ROC curve for PCT,
 \mathcal{H}_1 : the area under ROC curve for WBC \neq the area under the ROC curve for PCT.

Dependent ROC curves - comparison	
Analysis time	0,47sec.
Analysed variables	WBC;PCT;bacteremia
Count of missing data	7
Significance level	0,05
Grouping variable	bacteremia
Size	129
Size STATE + (yes)	32
Size STATE - (no)	97
Variable WBC	
Direction of diagnostic variable	stimulant
AUC	0,86130799
SE(AUC)	0,051727687
-95% CI	0,759923577
+95% CI	0,96269238
Variable PCT	
Direction of diagnostic variable	stimulant
AUC	0,895618557
SE(AUC)	0,049011126
-95% CI	0,79955852
+95% CI	0,991678596
DeLong's method	
AUC1-AUC2	0,034310567
SE(AUC1-AUC2)	0,022679669
-95% CI	0
+95% CI	0,078761905
Z statistic	1,512833666
p-value	0,130321915

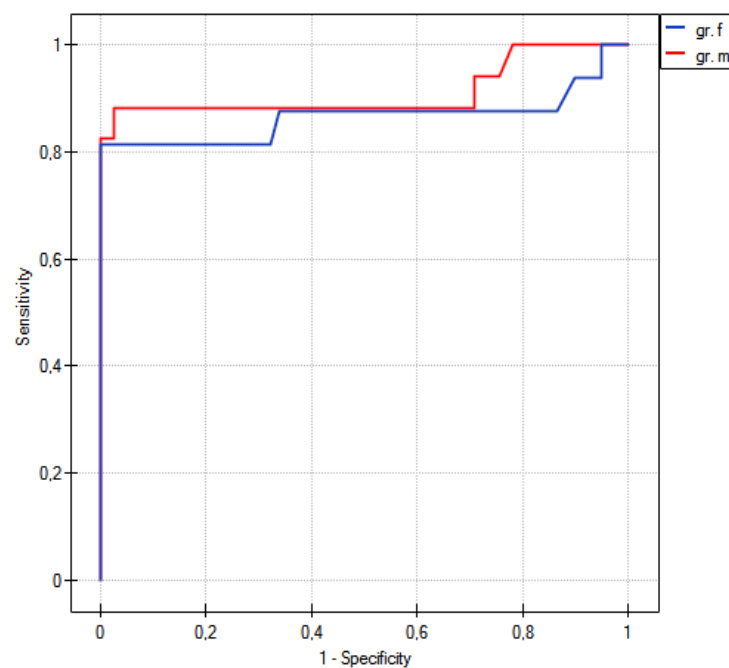


The calculated areas are $AUC_{WBC} = 0.8613$, $AUC_{PCT} = 0.8956$. On the basis of the adopted level $\alpha = 0.05$, based on the obtained value $p=0.13032$ we conclude that we cannot determine which of the parameters: WBC or PCT is better for diagnosing bacteremia.

ad2) PCT parameter is a stimulant (its value is high in bacteremia). In the course of the comparison of its diagnostic value for girls and boys we verify the following hypotheses:

- \mathcal{H}_0 : the area under ROC curve for PCT_f = the area under ROC curve for PCT_m ,
- \mathcal{H}_1 : the area under ROC curve for $PCT_f \neq$ the area under ROC curve for PCT_m .

Independent ROC curves - comparison	
Analysis time	0,36sec.
Analysed variables	PCT;bacteremia
Count of missing data	2
Significance level	0,05
Grouping variable	sex(f;m)
Direction of diagnostic variable	stimulant
Method	DeLong
Group name	f
Size	75
Size STATE + (yes)	16
Size STATE - (no)	59
AUC	0,864936441
SE(AUC)	0,079165996
-95% CI	0,709773958
+95% CI	1
Group name	m
Size	58
Size STATE + (yes)	17
Size STATE - (no)	41
AUC	0,911764706
SE(AUC)	0,059920399
-95% CI	0,794322908
+95% CI	1
AUC1-AUC2	0,046828265
SE(AUC1-AUC2)	0,099285997
Z statistic	0,47165025
p-value	0,637176453



The calculated areas are $AUC_f = 0.8649$, $AUC_m = 0.9118$. Therefore, on the basis of the adopted level $\alpha = 0.05$, based on the obtained value $p=0.6372$ we conclude that we cannot select the sex for which PCT parameter is better for diagnosing bacteremia.

17 MULTIDIMENSIONAL MODELS

17.1 PREPARATION OF THE VARIABLES FOR THE ANALYSIS IN MULTIDIMENSIONAL MODELS

17.1.1 Variable coding in multidimensional models

When preparing data for a multidimensional analysis there is the problem of appropriate coding of nominal and ordinal variables. That is an important element of preparing data for analysis as it is a key factor in the interpretation of the coefficients of a model. The nominal or ordinal variables divide the analyzed objects into two or more categories. The dichotomous variables (in two categories, $k = 2$) must only be appropriately coded, whereas the variables with many categories ($k > 2$) ought to be divided into dummy variables with two categories and coded.

$k = 2$ If a variable is dichotomous, it is the decision of the researcher how the data representing the variable will be entered, so any numerical codes can be entered, e.g. 0 and 1. In the program one can change one's coding into effect coding by selecting that option in the window of the selected multidimensional analysis. Such coding causes a replacement of the smaller value with value -1 and of the greater value with value 1.

$k > 2$ If a variable has many categories then in the window of the selected multidimensional analysis we select the button Dummy variables and set the reference/base category for those variables which we want to break into dummy variables. The variables will be dummy coded unless the effect coding option will be selected in the window of the analysis – in such a case, they will be coded as -1, 0, and 1.

Dummy coding is employed in order to answer, with the use of multidimensional models, the question: How do the (Y) results in any analyzed category differ from the results of the reference category. The coding consists in ascribing value 0 or 1 to each category of the given variable. The category coded as 0 is, then, the **reference category**.

$k = 2$ If the coded variable is dichotomous, then by placing it in a regression model we will obtain the coefficient calculated for it, (b_i). The coefficient is the reference of the value of the dependent variable Y for category 1 to the reference category (corrected with the remaining variables in the model).

$k > 2$ If the analyzed variable has more than two categories, then k categories are represented by $k - 1$ dummy variables with dummy coding. When creating variables with dummy coding one selects a category for which no dummy category is created. That category is treated as a reference category (as the value of each variable coded in the dummy coding is equal to 0. [0.2cm] When the X_1, X_2, \dots, X_{k-1} variables obtained in that way, with dummy coding, are placed in a regression model, then their b_1, b_2, \dots, b_{k-1} coefficients will be calculated.

b_1 is the reference of the Y results (for codes 1 in X_1) to the reference category (corrected with the remaining variables in the model);

b_2 is the reference of the Y results (for codes 1 in X_2) to the reference category (corrected with the remaining variables in the model);

...

b_{k-1} is the reference of the Y results (for codes 1 in X_{k-1}) to the reference category (corrected with the remaining variables in the model);

Example

We code, in accordance with dummy coding, the sex variable with two categories (the male sex will be selected as the reference category), and the education variable with 4 categories (elementary education will be selected as the reference category).

Sex	Coded sex	Education	Coded education		
			vocational	secondary	tertiary
f	1	elementary	0	0	0
f	1	elementary	0	0	0
f	1	elementary	0	0	0
m	0	vocational	1	0	0
m	0	vocational	1	0	0
f	1	vocational	1	0	0
f	1	vocational	1	0	0
m	0	secondary	0	1	0
m	0	secondary	0	1	0
f	1	secondary	0	1	0
m	0	secondary	0	1	0
f	1	tertiary	0	0	1
m	0	tertiary	0	0	1
f	1	tertiary	0	0	1
m	0	tertiary	0	0	1
m	0	tertiary	0	0	1
...

Building on the basis of dummy variables, in a multiple regression model, we might want to check what impact the variables have on a dependent variable, e.g. Y = the amount of earnings (in thousands of PLN). As a result of such an analysis we will obtain sample coefficients for each dummy variable:

- for sex the statistically significant coefficient $b_i = -0.5$, which means that average women's wages are a half of a thousand PLN lower than men's wages, assuming that all other variables in the model remain unchanged;
- for vocational education the statistically significant coefficient $b_i = 0.6$, which means that the average wages of people with elementary education are 0.6 of a thousand PLN higher than those of people with elementary education, assuming that all other variables in the model remain unchanged;
- for secondary education the statistically significant coefficient $b_i = 1$, which means that the average wages of people with secondary education are a thousand PLN higher than those of people with elementary education, assuming that all other variables in the model remain unchanged;
- for tertiary-level education the statistically significant coefficient $b_i = 1.5$, which means that the average wages of people with tertiary-level education are 1.5 PLN higher than those of people with elementary education, assuming that all other variables in the model remain unchanged;

Effect coding is used to answer, with the use of multidimensional models, the question: How do (Y) results in each analyzed category differ from the results of the (unweighted) mean obtained from the sample. The coding consists in ascribing value -1 or 1 to each category of the given variable. The category coded as -1 is then the **base category**

$k = 2$ If the coded variable is dichotomous, then by placing it in a regression model we will obtain the coefficient calculated for it, (b_i). The coefficient is the reference of Y for category 1 to the unweighted general mean (corrected with the remaining variables in the model).

If the analyzed variable has more than two categories, then k categories are represented by $k - 1$ dummy variables with effect coding. When creating variables with effect coding a category is selected for which no separate variable is made. The category is treated in the models as a base category (as in each variable made by effect coding it has values -1).

When the X_1, X_2, \dots, X_{k-1} variables obtained in that way, with effect coding, are placed in

a regression model, then their b_1, b_2, \dots, b_{k-1} coefficients will be calculated.

b_1 is the reference of the Y results (for codes 1 in X_1) to the unweighted general mean (corrected by the remaining variables in the model);

b_2 is the reference of the Y results (for codes 1 in X_2) to the unweighted general mean (corrected by the remaining variables in the model);

...

b_{k-1} is the reference of the Y results (for codes 1 in X_{k-1}) to the unweighted general mean (corrected by the remaining variables in the model);

Example

With the use of effect coding we will code the sex variable with two categories (the male category will be the base category) and a variable informing about the region of residence in the analyzed country. 5 regions were selected: northern, southern, eastern, western, and central. The central region will be the base one.

Sex	Coded sex	Regions of residence	Coded regions			
			western	eastern	northern	southern
f	1	central	-1	-1	-1	-1
f	1	central	-1	-1	-1	-1
f	1	central	-1	-1	-1	-1
m	-1	western	1	0	0	0
m	-1	western	1	0	0	0
f	1	western	1	0	0	0
f	1	western	1	0	0	0
m	-1	eastern	0	1	0	0
m	-1	eastern	0	1	0	0
f	1	eastern	0	1	0	0
m	-1	eastern	0	1	0	0
f	1	northern	0	0	1	0
m	-1	northern	0	0	1	0
f	1	southern	0	0	0	1
m	-1	southern	0	0	0	1
m	-1	southern	0	0	0	1
...

Building on the basis of dummy variables, in a multiple regression model, we might want to check what impact the variables have on a dependent variable, e.g. Y = the amount of earnings (expressed in thousands of PLN). As a result of such an analysis we will obtain sample coefficients for each dummy variable:

- for sex the statistically significant coefficient $b_i = -0.5$, which means that the average women's wages are a half of a thousand PLN lower than the average wages in the country, assuming that the other variables in the model remain unchanged;
- for the western region the statistically significant coefficient $b_i = 0.6$, which means that the average wages of people living in the western region of the country are 0.6 thousand PLN higher than the average wages in the country, assuming that the other variables in the model remain unchanged;
- for the eastern region the statistically significant coefficient $b_i = -1$, which means that the average wages of people living in the eastern region of the country are a thousand PLN lower than the average wages in the country, assuming that the other variables in the model remain unchanged;
- for the northern region the statistically significant coefficient $b_i = 0.4$, which means that the

average wages of people living in the western region of the country are 0.4 thousand PLN higher than the average wages in the country, assuming that the other variables in the model remain unchanged;

- for the southern region the statistically significant coefficient $b_i = 0.1$, which means that the average wages of people living in the southern region of the country do not differ in a statistically significant manner from the average wages in the country, assuming that the other variables in the model remain unchanged;

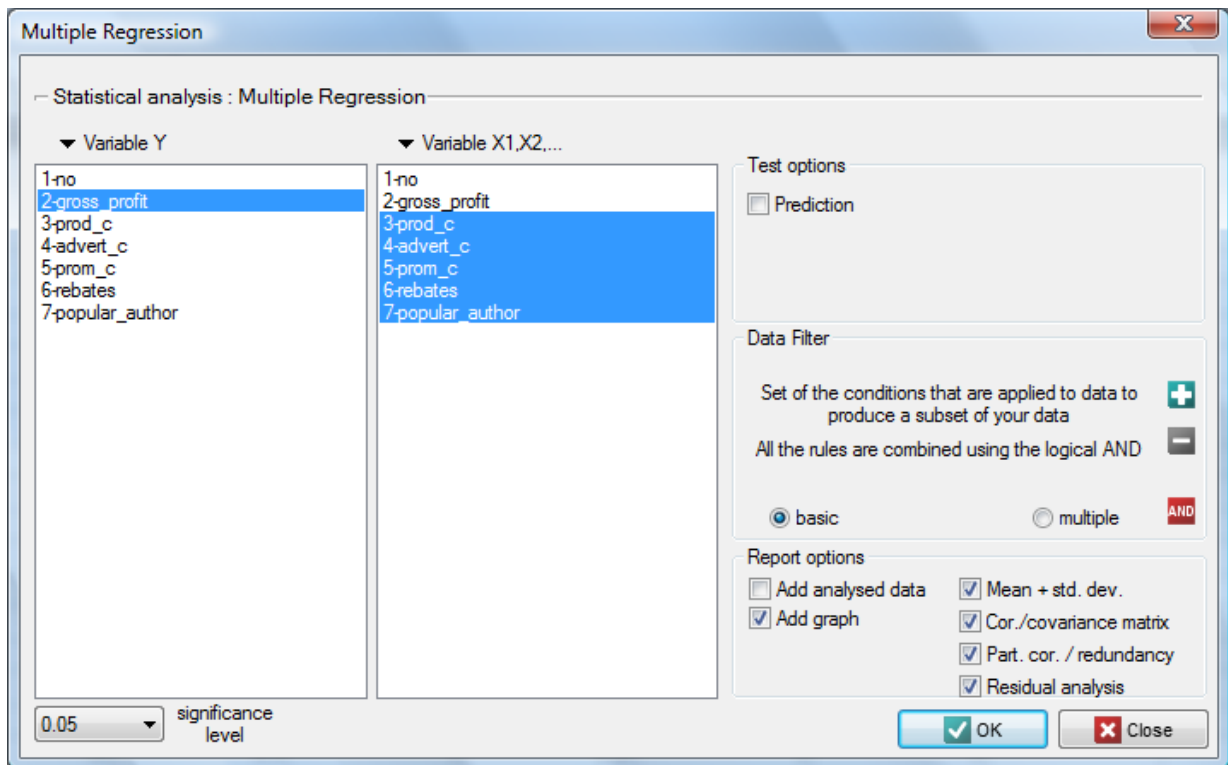
17.1.2 Interactions

Interactions are considered in multidimensional models. Their presence means that the influence of the independent variable (X_1) on the dependent variable (Y) differs depending on the level of another independent variable (X_2) or a series of other independent variables. To discuss the interactions in multidimensional models one must determine the variables informing about possible interactions, i.e. the product of appropriate variables. For that purpose we select the Interactions button in the window of the selected multidimensional analysis. In the window of interactions settings, with the CTRL button pressed, we determine the variables which are to form interactions and transfer the variables into the neighboring list with the use of an arrow. By pressing the OK button we will obtain appropriate columns in the datasheet.

In the analysis of the interaction the choice of appropriate coding of dichotomous variables allows the avoidance of the over-parametrization related to interactions. Over-parametrization causes the effects of the lower order for dichotomous variables to be redundant with respect to the confounding interactions of the higher order. As a result, the inclusion of the interactions of the higher order in the model annuls the effect of the interactions of the lower orders, not allowing an appropriate evaluation of the latter. In order to avoid the over-parametrization in a model in which there are interactions of dichotomous variables it is recommended to choose the option effect coding.

17.2 MULTIPLE LINEAR REGRESSION

The window with settings for Multiple Regression is accessed via the menu Statistics → Multidimensional Models → Multiple Regression



The constructed model of linear regression allows the study of the influence of many independent variables (X_1, X_2, \dots, X_k) on one dependent variable (Y). The most frequently used variety of multiple regression is Multiple Linear Regression. It is an extension of linear regression models based on [Pearson's linear correlation coefficient](#). It presumes the existence of a linear relation between the studied variables. The linear model of multiple regression has the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon.$$

where:

Y - dependent variable, explained by the model,
 X_1, X_2, \dots, X_k - independent variables, explanatory,
 $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ - parameters,
 ϵ - random parameter (model residual).

If the model was created on the basis of a data sample of size n the above equation can be presented in the form of a matrix:

$$Y = X\beta + \epsilon.$$

where:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

In such a case, the solution of the equation is the vector of the estimates of parameters $\beta_0, \beta_1, \dots, \beta_k$ called **regression coefficients**:

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}.$$

Those coefficients are estimated with the help of the classical **least squares method**. On the basis of those values we can infer the magnitude of the effect of the independent variable (for which the coefficient was estimated) on the dependent variable. They inform by how many units will the dependent variable change when the independent variable is changed by 1 unit. There is a certain error of estimation for each coefficient. The magnitude of that error is estimated from the following formula:

$$SE_b = \sqrt{\frac{1}{n - (k + 1)} e^T e (X^T X)^{-1}},$$

where:

$e = Y - \hat{Y}$ is the vector of **model residuals** (the difference between the actual values of the dependent variable Y and the values \hat{Y} predicted on the basis of the model).

Note

When constructing the model one should remember that the number of observations has to be greater than or equal to the number of the estimated parameters of the model ($n \geq k + 1$).

17.2.1 Model verification

- **Statistical significance of particular variables in the model.**

On the basis of the coefficient and its error of estimation we can infer if the independent variable for which the coefficient was estimated has a significant effect on the dependent variable. For that purpose we use t-test.

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : & \beta_i = 0, \\ \mathcal{H}_1 : & \beta_i \neq 0. \end{aligned}$$

Let us estimate the test statistics according to the formula below:

$$t = \frac{b_i}{SE_{b_i}}$$

The test statistics has **t-Student distribution** with $n - k$ degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha & \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

- **The quality of the constructed model** of multiple linear regression can be evaluated with the help of several measures.

– **The standard error of estimation** – it is the measure of model adequacy:

$$SE_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n - (k + 1)}}.$$

The measure is based on model residuals $e_i = y_i - \hat{y}_i$, that is on the discrepancy between the actual values of the dependent variable y_i in the sample and the values of the independent variable \hat{y}_i estimated on the basis of the constructed model. It would be best if the difference were as close to zero as possible for all studied properties of the sample. Therefore, for the model to be well-fitting, the standard error of estimation (SE_e), expressed as e_i variance, should be the smallest possible.

- **Multiple correlation coefficient** $R = \sqrt{R^2} \in < 0; 1 >$ – defines the strength of the effect of the set of variables X_1, X_2, \dots, X_k on the dependent variable Y .
- **Multiple determination coefficient** R^2 – it is the measure of model adequacy.

The value of that coefficient falls within the range of $< 0; 1 >$, where 1 means excellent model adequacy, 0 – a complete lack of adequacy. The estimation is made using the following formula:

$$T_{SS} = E_{SS} + R_{SS},$$

where:

T_{SS} – total sum of squares,

E_{SS} – the sum of squares explained by the model,

R_{SS} – residual sum of squares.

The coefficient of determination is estimated from the formula:

$$R^2 = \frac{T_{SS}}{E_{SS}}.$$

It expresses the percentage of the variability of the dependent variable explained by the model.

As the value of the coefficient R^2 depends on model adequacy but is also influenced by the number of variables in the model and by the sample size, there are situations in which it can be encumbered with a certain error. That is why a corrected value of that parameter is estimated:

$$R_{adj}^2 = R^2 - \frac{k(1 - R^2)}{n - (k + 1)}.$$

- **Statistical significance of all variables in the model**

The basic tool for the evaluation of the significance of all variables in the model is the analysis of variance test (the F-test). The test simultaneously verifies 3 equivalent hypotheses:

$$\begin{array}{ll} \mathcal{H}_0 : & \text{all } \beta_i = 0, & \mathcal{H}_1 : & \text{exists } \beta_i \neq 0; \\ \mathcal{H}_0 : & R^2 = 0, & \mathcal{H}_1 : & R^2 \neq 0; \\ \mathcal{H}_0 : & \text{linearity of the relation,} & \mathcal{H}_1 : & \text{a lack of a linear relation.} \end{array}$$

The test statistics has the form presented below:

$$F = \frac{E_{MS}}{R_{MS}}$$

where:

$E_{MS} = \frac{E_{SS}}{df_E}$ – the mean square explained by the model,

$R_{MS} = \frac{R_{SS}}{df_R}$ – residual mean square,

$df_E = k, df_R = n - (k + 1)$ – appropriate degrees of freedom.

That statistics is subject to **F-Snedecor distribution** with df_E and df_R degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{array}{ll} \text{if } p \leq \alpha & \implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

17.2.2 More information about the variables in the model

- **Standardized** b_1, b_2, \dots, b_k — In contrast to raw parameters (which are expressed in different units of measure, depending on the described variable, and are not directly comparable) the standardized estimates of the parameters of the model allow the comparison of the contribution of particular variables to the explanation of the variance of the dependent variable Y .
- **Correlation matrix** — contains information about the strength of the relation between particular variables, that is the **Pearson's correlation** coefficient $r_p \in < -1; 1 >$. The coefficient is used for the study of the correlation of each pair of variables, without taking into consideration the effect of the remaining variables in the model.
- **Covariance matrix** — similarly to the correlation matrix it contains information about the linear relation among particular variables. That value is not standardized.
- **Partial correlation coefficient** — falls within the range $< -1; 1 >$ and is the measure of correlation between the specific independent variable X_i (taking into account its correlation with the remaining variables in the model) and the dependent variable Y (taking into account its correlation with the remaining variables in the model).
The square of that coefficient is the **partial determination coefficient** — it falls within the range $< 0; 1 >$ and defines the relation of only the variance of the given independent variable X_i with that variance of the dependent variable Y which was not explained by other variables in the model.
The closer the value of those coefficients to 0, the more useless the information carried by the studied variable, which means the variable is superfluous.
- **Semipartial correlation coefficient** — falls within the range $< -1; 1 >$ and is the measure of correlation between the specific independent variable X_i (taking into account its correlation with the remaining variables in the model) and the dependent variable Y (NOT taking into account its correlation with the remaining variables in the model).
The square of that coefficient is the **semipartial determination coefficient** — it falls within the range $< 0; 1 >$ and defines the relation of only the variance of the given independent variable X_i with the complete variance of the dependent variable Y .
The closer the value of those coefficients to 0, the more useless the information carried by the studied variable, which means the variable is superfluous.
- **R-squared** ($R^2 \in < 0; 1 >$) - it represents the percentage of variance of the given independent variable X_i , explained by the remaining independent variables. The closer to value 1 the stronger the linear relation of the studied variable with the remaining independent variables, which can mean that the variable is a superfluous one.
- **Tolerance** $= 1 - R^2 \in < 0; 1 >$ — it represents the percentage of variance of the given independent variable X_i , NOT explained by the remaining independent variables. The closer the value of tolerance is to 0 the stronger the linear relation of the studied variable with the remaining independent variables, which can mean that the variable is a superfluous one.
- **A comparison of a full model with a model in which a given variable is removed**
The comparison of the two model is made with by means of:
 - **F test**, in a situation in which one variable or more are removed from the model (see: the comparison of models),
 - **t-test**, when only one variable is removed from the model. It is the same test that is used for studying the significance of particular variables in the model.

In the case of removing only one variable the results of both tests are identical.

If the difference between the compared models is statistically significant (the value $p \leq \alpha$), the full model is significantly better than the reduced model. It means that the studied variable is not superfluous, it has a significant effect on the given model and should not be removed from it.

- **Scatter plots**

The charts allow a subjective evaluation of linearity of the relation among the variables and an identification of outliers. Additionally, scatter plots can be useful in an analysis of model residuals.

17.2.3 Analysis of model residuals

To obtain a correct regression model we should check the basic assumptions concerning model residuals.

- **Outliers**

The study of the model residual can be a quick source of knowledge about outlier values. Such observations can disturb the equation of the regression to a large extent because they have a great effect on the values of the coefficients in the equation. If the given residual e_i deviates by more than 3 standard deviations from the mean value, such an observation can be classified as an outlier. A removal of an outlier can greatly enhance the model.

- **Normality of distribution of model residuals**

The assumption is checked with the help of [Lilliefors test](#). A big difference between the residuals distribution and the normal distribution (the value $p \leq \alpha$) can impair the evaluation of the significance of the coefficients of particular variables in the model.

- **Homoscedasticity (homogeneity of variance)**

To check if there are areas in which the variance of model residuals is increased or decreased we use the charts of:

- the residual with respect to predicted values
- the square of the residual with respect to predicted values
- the residual with respect to observed values
- the square of the residual with respect to observed values

- **Autocorrelation of model residuals**

For the constructed model to be deemed correct the values of residuals should not be correlated with one another (for all pairs e_i, e_j). The assumption can be checked by computing the Durbin-Watson statistic.

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2},$$

To test for positive autocorrelation on the significance level α we check the position of the statistics d with respect to the upper ($d_{U,\alpha}$) and lower ($d_{L,\alpha}$) critical value:

- If $d < d_{L,\alpha}$ – the errors are positively correlated;
- If $d > d_{U,\alpha}$ – the errors are not positively correlated;
- If $d_{L,\alpha} < d < d_{U,\alpha}$ – the test result is ambiguous.

To test for negative autocorrelation on the significance level α we check the position of the value $4 - d$ with respect to the upper ($d_{U,\alpha}$) and lower ($d_{L,\alpha}$) critical value:

- If $4 - d < d_{L,\alpha}$ – the errors are negatively correlated;

- If $4 - d > d_{U,\alpha}$ – the errors are not negatively correlated;
- If $d_{L,\alpha} < 4 - d < d_{U,\alpha}$ – the test result is ambiguous.

The critical values of the Durbin-Watson test for the significance level $\alpha = 0.05$ are on the website www.pqstat.com – the source of the: Savin and White tables (1977)[74]

17.2.4 Prediction on the basis of the model

Most often, the last stage of regression analysis is the use of the constructed and verified model for prediction. Predicting the value of the dependent variable is possible for the studied values of independent variables. The computed value is estimated with a certain error. That is why, additionally, limits resulting from error are estimated for the estimated value:

- for the expected value, confidence limits are estimated,
- for a single point, prediction limits are estimated.

EXAMPLE 17.1. (publisher.pqs file)

A certain book publisher wanted to learn how was gross profit from sales influenced by such variables as: production cost, advertising costs, direct promotion cost, the sum of discounts made, and the author's popularity. For that purpose he analyzed 40 titles published during the previous year. A part of the data is presented in the image below:

no	gross_profit	prod_c	advert_c	prom_c	rebates	popular_author
1	58	7.9	9	0.38	1.8	1
2	63	10.1	10	0.59	2.4	0
3	27	3	7	0.7	1.7	0
4	35	6	3	0.21	2.6	1
5	34	6.6	2.1	0.13	2.2	0
6	48	10.7	1	0.08	2.1	1
7	14	2.7	0.7	0.06	0.3	0
8	63.5	12	5	0.56	1.7	0

The first five variables are expressed in thousands of dollars - so they are variables gathered on an interval scale. The last variable: the author's popularity – is a dichotomic variable, where 1 stands for a known author, and 0 stands for an unknown author.

On the basis of the knowledge gained from the analysis the publisher wants to predict the gross profit from the next published book written by a known author. The expenses the publisher will bear are: production cost ≈ 11 , advertising costs ≈ 13 , direct promotion costs ≈ 0.5 , the sum of discounts made ≈ 0.5 .

We construct the model of multiple linear regression, selecting: gross profit – as the dependent variable Y , production cost, advertising costs, direct promotion costs, the sum of discounts made, the author's popularity – as the independent variables X_1, X_2, X_3, X_4, X_5 . As a result, the coefficients of the regression equation will be estimated, together with measures which will allow the evaluation of the quality of the model.

Multiple Regression	
Analysis time	0.52sec.
Analysed variables	gross_profit;prod_c;advert_c
Significance level	0.05
Size	40
Number of estimated parameters	6
R	0.922483
R2	0.850974
Adjusted R2	0.829059
Standard error of estimation	8.086501
Residual sum of squares	2223.31121
Total sum of squares	14918.99775
Explained sum of squares	12695.68654
F	38.829772
p-value	<0.000001

Model	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. er	mean	stand. dev
intercept	4.175186	4.772779	-5.524268	13.874639	0.874791	0.387825				
prod_c	2.560709	0.501507	1.541525	3.579894	5.106033	0.000013	0.422818	0.082807	7.4675	3.229463
advert_c	1.998235	0.359065	1.268527	2.727943	5.565106	0.000003	0.461287	0.082889	5.85	4.515046
prom_c	4.668238	4.791644	-5.069555	14.40603	0.974245	0.336816	0.066242	0.067993	0.5145	0.277534
rebates	1.423171	1.404811	-1.431749	4.278091	1.013069	0.318183	0.067478	0.066608	1.605	0.927348
popular_author	10.153717	2.782567	4.498861	15.808574	3.649047	0.000874	0.261561	0.071679	0.55	0.503831

On the basis of the estimated value of the coefficient b , the relationship between gross profit and all independent variables can be described by means of the equation:

$$profit_{gross} = 4.18 + 2.56(c_{prod}) + 2(c_{adv}) + 4.67(c_{prom}) + 1.42(discounts) + 10.15(popul_{author}) + [8.09]$$

The obtained coefficients are interpreted in the following manner:

- If the production cost increases by 1 thousand dollars, then gross profit will increase by about 2.56 thousand dollars, assuming that the remaining variables do not change;
- If the production cost increases by 1 thousand dollars, then gross profit will increase by about 2 thousand dollars, assuming that the remaining variables do not change;
- If the production cost increases by 1 thousand dollars, then gross profit will increase by about 4.67 thousand dollars, assuming that the remaining variables do not change;
- If the sum of the discounts made increases by 1 thousand dollars, then gross profit will increase by about 1.42 thousand dollars, assuming that the remaining variables do not change;
- If the book has been written by a known author (marked as 1), then in the model the author's popularity is assumed to be the value 1 and we get the equation:

$$profit_{gross} = 14.33 + 2.56(c_{prod}) + 2(c_{adv}) + 4.67(c_{prom}) + 1.42(discounts)$$

If the book has been written by an unknown author (marked as 0), then in the model the author's popularity is assumed to be the value 0 and we get the equation:

$$profit_{gross} = 4.18 + 2.56(c_{prod}) + 2(c_{adv}) + 4.67(c_{prom}) + 1.42(discounts)$$

The result of t-test for each variable shows that only the production cost, advertising costs, and author's popularity have a significant influence on the profit gained. At the same time, that standardized coefficients b are the greatest for those variables.

Additionally, the model is very well-fitting, which is confirmed by: the small standard error of estimation $SE_e = 8.086501$, the high value of the multiple determination coefficient $R^2 = 0.850974$, the corrected multiple determination coefficient $R^2_{adj} = 0.829059$, and the result of the F-test of variance analysis: $p < 0.000001$.

On the basis of the interpretation of the results obtained so far we can assume that a part of the variables does not have a significant effect on the profit and can be superfluous.

For the model to be well formulated the interval independent variables ought to be strongly correlated with the dependent variable and be relatively weakly correlated with one another. That can be checked by computing the correlation matrix and the covariance matrix:

Correlations						
	gross_prof	prod_c	advert_c	prom_c	rebates	popular_ai
gross_profit	1	0.770685	0.794843	0.071095	0.131924	0.553803
prod_c	0.770685	1	0.558843	-0.079792	0.092951	0.340624
advert_c	0.794843	0.558843	1	0.119889	0.056708	0.326878
prom_c	0.071095	-0.079792	0.119889	1	-0.056478	-0.049327
rebates	0.131924	0.092951	0.056708	-0.056478	1	0.010427
popular_author	0.553803	0.340624	0.326878	-0.049327	0.010427	1

Covariance						
	gross_prof	prod_c	advert_c	prom_c	rebates	popular_ai
gross_profit	382.53840	48.679353	70.190897	0.385914	2.392782	5.457308
prod_c	48.679353	10.429429	8.14859	-0.071517	0.278372	0.554231
advert_c	70.190897	8.14859	20.385641	0.150231	0.237436	0.74359
prom_c	0.385914	-0.071517	0.150231	0.077025	-0.014536	-0.006897
rebates	2.392782	0.278372	0.237436	-0.014536	0.859974	0.004872
popular_author	5.457308	0.554231	0.74359	-0.006897	0.004872	0.253846

The most coherent information which allows finding those variables in the model which are superfluous is given by the partial and semipartial correlation analysis as well as redundancy analysis:

Part. Semipart. Cor.						
	partial	semipartia	tolerance	R^2	t stat.	p-value
prod_c	0.658793	-0.338045	0.639209	0.360791	5.106033	0.000013
advert_c	0.690424	-0.368438	0.637949	0.362051	5.565106	0.000003
prom_c	0.164797	-0.0645	0.948099	0.051901	0.974245	0.336816
rebates	0.171176	-0.06707	0.987951	0.012049	1.013069	0.318183
popular_author	0.53049	-0.241585	0.85309	0.14691	3.649047	0.000874

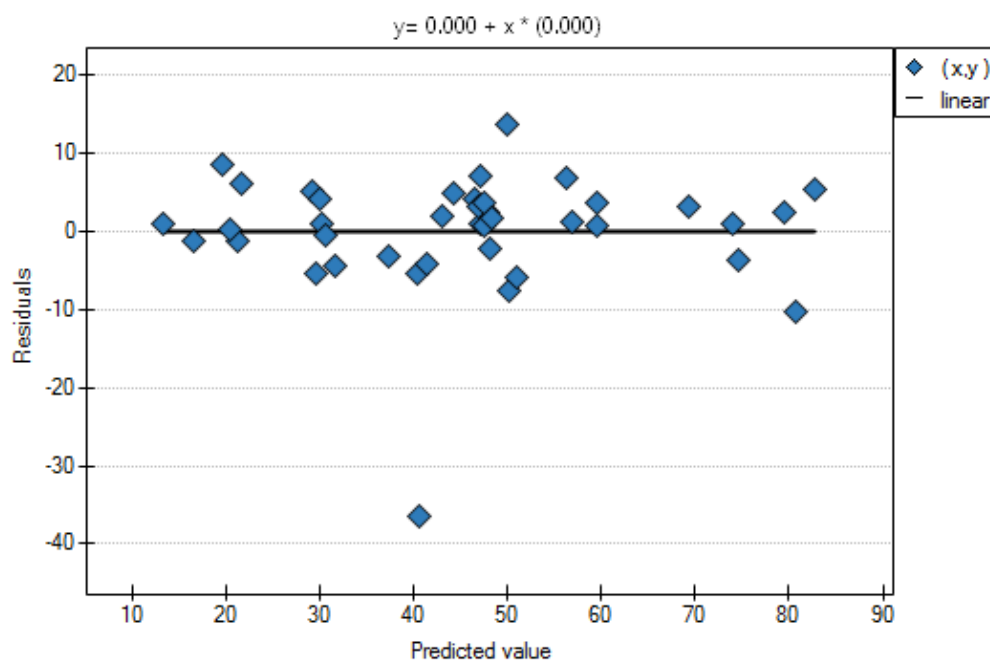
The values of coefficients of partial and semipartial correlation indicate that the smallest contribution into the constructed model is that of direct promotion costs and the sum of discounts made. However, those variables are the least correlated with model residuals, which is indicated by the low value R^2 and the high tolerance value. All in all, from the statistical point of view, models without those variables would not be worse than the current model (see the result of t-test for model comparison). The decision

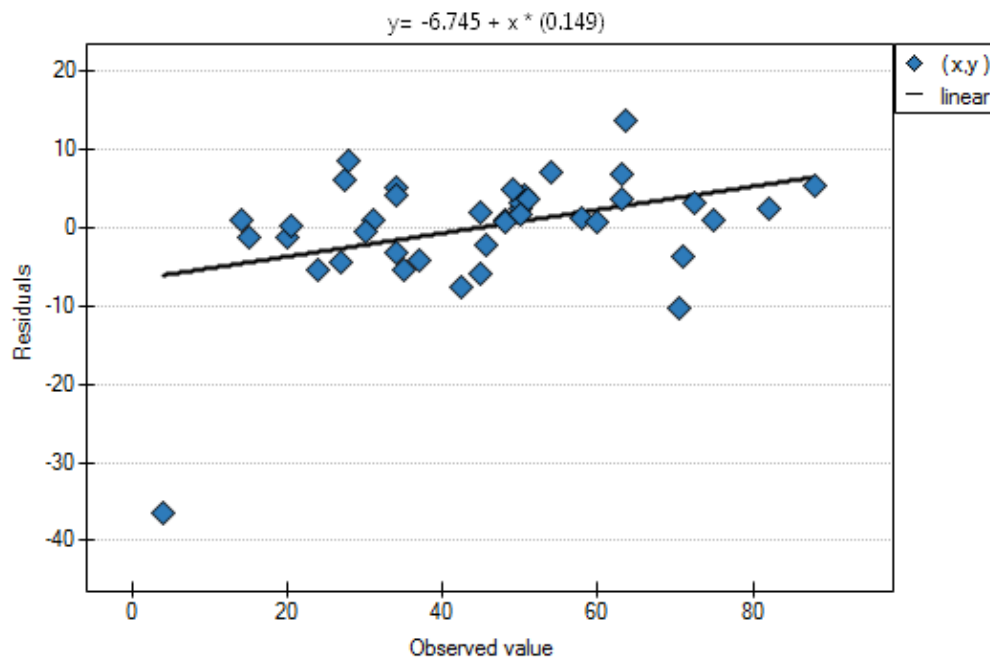
about whether or not to leave that model or to construct a new one without the direct promotion costs and the sum of discounts made, belongs to the researcher. We will leave the current model.

Finally, we will analyze the residuals. A part of that analysis is presented below:

Residual analysis										
	predicted \	residual	standard r	<=-3sd	(-3sd;2sd]	(-2sd;sd]	(-sd;sd)	[sd;2sd)	[2sd;3sd)	>=3sd
1	56.87826	1.12174	0.138718				*			
2	56.190571	6.809429	0.842074				*			
3	31.532115	-4.532115	-0.560454				*			
4	40.368438	-5.368438	-0.663877				*			
5	29.010008	4.989992	0.617077				*			
6	47.088847	0.911153	0.112676				*			
7	13.194911	0.805089	0.09956				*			
8	49.928477	13.571523	1.678293					*		
9	46.501478	3.998522	0.494469				*			
10	47.776747	2.223253	0.274934				*			
11	47.989623	-2.289623	-0.283141				*			
12	79.536147	2.463853	0.304687				*			
13	29.921482	4.078518	0.504361				*			
14	30.206017	0.793983	0.098186				*			
15	46.840918	3.159082	0.390661				*			
16	40.467863	-36.467863	-4.509721	*						
17	47.110045	6.889955	0.852032				*			
18	50.197864	-7.697864	-0.95194			*				
19	43.035765	1.964235	0.242903				*			
20	44.211462	4.788538	0.592164				*			
21	29.452408	-5.452408	-0.67426				*			

It is noticeable that one of the model residuals is an outlier – it deviates by more than 3 standard deviations from the mean value. It is observation number 16. The observation can be easily found by drawing a chart of residuals with respect to observed or expected values of the variable Y .





That outlier undermines the assumption concerning homoscedasticity. The assumption of homoscedasticity would be confirmed (that is, residuals variance presented on the axis Y would be similar when we move along the axis X), if we rejected that point. Additionally, the distribution of residuals deviates slightly from normal distribution (the value p of Liliefors test is $p = 0.016415$):

normality of residuals	
d statistic	0.155181
degrees of freedom	40
p-value	0.016415

When we take a closer look of the outlier (position 16 in the data for the task) we see that the book is the only one for which the costs are higher than gross profit (gross profit=4 thousand dollars, the sum of costs = $(8+6+0.33+1.6) = 15.93$ thousand dollars).

The obtained model can be corrected by removing the outlier. For that purpose, another analysis has to be conducted, with a filter switched on which will exclude the outlier.

Data Filter		
variable	condition	value
1-no	<>	16

Buttons: +, -, AND

As a result, we receive a model which is very similar to the previous one but is encumbered with a smaller error and is more adequate:

Multiple Regression	
Analysis time	0.50sec.
Analysed variables	gross_profit;prod_c;advert_c
Significance level	0.05
Data Filter	no<>16
Size	39
Number of estimated parameters	6
R	0.969923
R2	0.940751
Adjusted R2	0.931774
Standard error of estimation	4.863281
Residual sum of squares	780.499712
Total sum of squares	13173.170769
Explained sum of squares	12392.671057
F	104.793926
p-value	<0.000001

Model	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. er	mean	stand. dev
intercept	6.892628	2.891394	1.010044	12.775213	2.383843	0.023041				
prod_c	2.678933	0.301989	2.064531	3.293335	8.870952	<0.000001	0.47057	0.053046	7.453846	3.27051
advert_c	2.080912	0.216204	1.641042	2.520781	9.624775	<0.000001	0.511207	0.053114	5.846154	4.574002
prom_c	1.920214	2.903129	-3.986247	7.826675	0.661429	0.512929	0.028828	0.043584	0.519231	0.279524
rebates	1.325426	0.844957	-0.393651	3.044503	1.568632	0.126274	0.066878	0.042635	1.605128	0.93947
popular_author	7.382624	1.710653	3.902274	10.862973	4.315676	0.000136	0.199191	0.046155	0.564103	0.502356

$$profit_{gross} = 6.89 + 2.68(c_{prod}) + 2.08(c_{adv}) + 1.92(c_{prom}) + 1.33(discounts) + 7.38(popul_{author}) + [4.86]$$

The final version of the model will be used for prediction. On the basis of the predicted costs amounting to:

production cost \approx 11 thousand dollars,

advertising costs \approx 13 thousand dollars,

direct promotion costs \approx 0.5 thousand dollars,

the sum of discounts made \approx 0.5 thousand dollars,

and the fact that the author is known (the author's popularity \approx 1) we calculate the predicted gross profit together with the confidence interval:

Prediction for 3-prod_c	11
Prediction for 4-advert_c	13
Prediction for 5-prom_c	0.5
Prediction for 6-rebates	0.5
Prediction for 7-popular_author	1
Prediction of Y value	72.418189
-95% CI (for point)	61.856275
+95% CI (for point)	82.980103
-95% CI (for expected values)	68.722994
+95% CI (for expected values)	76.113384

The predicted profit is 72 thousand dollars.

Note

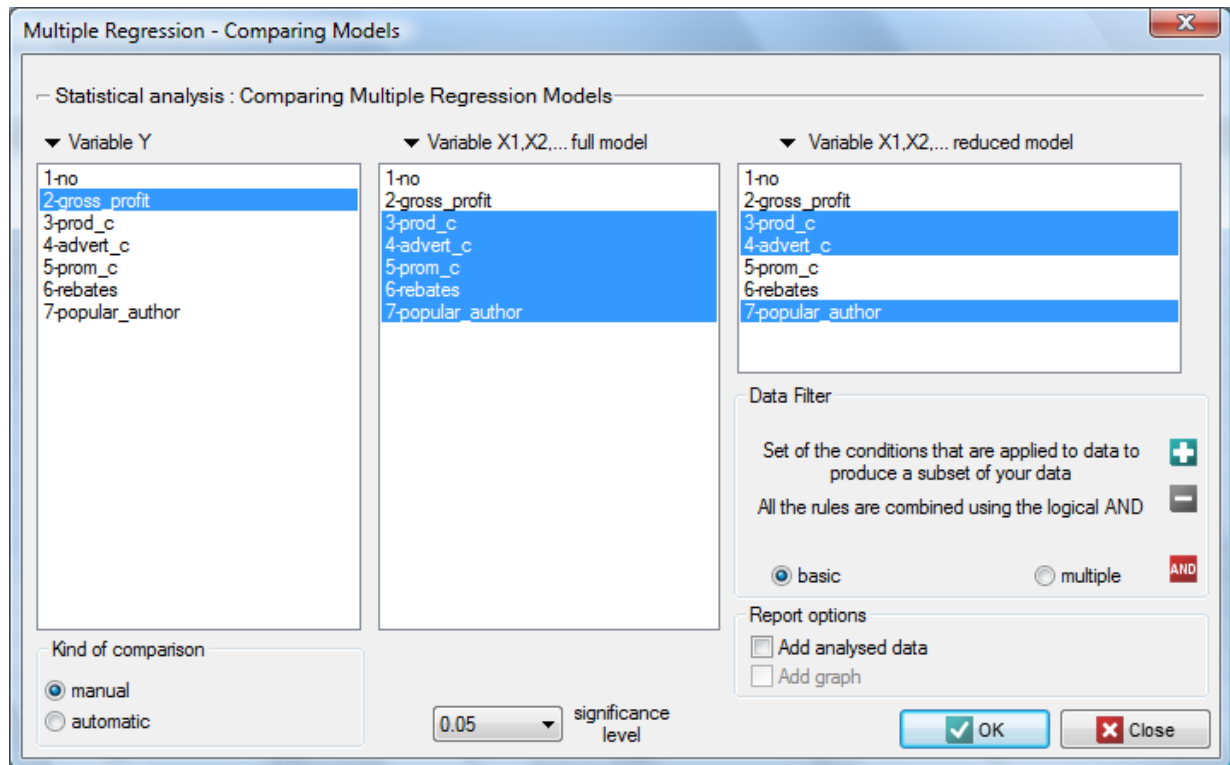
To be able to consider the nominal independent variable in many categories in the model, the variable ought to be decomposed into several dummy variables, in 2 categories, before the analysis

Note

To take into consideration the interactions of independent variables, a variable which is the result of multiplying the variables participating in the interaction ought to be introduced into the model.

17.3 COMPARISON OF MULTIPLE LINEAR REGRESSION MODELS

The window with settings for model comparison is accessed via the menu Statistics → Multidimensional models → Multiple regression — model comparison



The multiple linear regression offers the possibility of simultaneous analysis of many independent variables. There appears, then, the problem of choosing the optimum model. Too large a model involves a plethora of information in which the important ones may get lost. Too small a model involves the risk of omitting those features which could describe the studied phenomenon in a reliable manner. Because it is not the number of variables in the model but their quality that determines the quality of the model. To make a proper selection of independent variables it is necessary to have knowledge and experience connected with the studied phenomenon. One has to remember to put into the model variables strongly correlated with the dependent variable and weakly correlated with one another.

There is no single, simple statistical rule which would decide about the number of variables necessary in the model. The measures of model adequacy most frequently used in a comparison are: R_{adj}^2 — the corrected value of multiple determination coefficient (the higher the value the more adequate the model), SE_e — the standard error of estimation (the lower the value the more adequate the model). For that purpose, the F-test based on the multiple determination coefficient R^2 can also be used. The test is used to verify the hypothesis that the adequacy of both compared models is equally good.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : R_{FM}^2 &= R_{RM}^2, \\ \mathcal{H}_1 : R_{FM}^2 &\neq R_{RM}^2,\end{aligned}$$

where:

R_{FM}^2, R_{RM}^2 — multiple determination coefficients in compared models (full and reduced).

The test statistics has the form presented below:

$$F = \frac{R_{FM}^2 - R_{RM}^2}{k_{FM} - k_{RM}} \cdot \frac{n - k_{FM} - 1}{1 - R_{FM}^2},$$

The statistics is subject to **F-Snedecor distribution** with $df_1 = k_{FM} - k_{RM}$ and $df_2 = n - k_{FM} - 1$ degrees of freedom.

The **p value**, designated on the basis of the **test statistic**, is compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

If the compared models do not differ significantly, we should select the one with a smaller number of variables. Because a lack of a difference means that the variables present in the full model but absent from the reduced model do not carry significant information. However, if the difference in the quality of model adequacy is statistically significant, it means that one of them (the one with the greater number of variables, with a greater R^2) is significantly better than the other one.

In the program PQStat the comparison of models can be done manually or automatically.

- **Manual** model comparison – construction of 2 models:
 - a full model – a model with a greater number of variables,
 - a reduced model – a model with a smaller number of variables – such a model is created from the full model by removing those variables which are superfluous from the perspective of studying a given phenomenon.

The choice of independent variables in the compared models and, subsequently, the choice of a better model on the basis of the results of the comparison, is made by the researcher.

- **Automatic** model comparison is done in several steps:
 - step 1 Constructing the model with the use of all variables.
 - step 2 Removing one variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.
 - step 3 A comparison of the full and the reduced model.
 - step 4 Removing another variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.
 - step 5 A comparison of the previous and the newly reduced model.
 - ...

In that way numerous, ever smaller models are created. The last model only contains 1 independent variable.

As a result, each model is described with the help of adequacy measures (R_{adj}^2, SE_e), and the subsequent (neighboring) models are compared by means of the F-test. The model which is finally marked as statistically best is the model with the greatest R_{adj}^2 and the smallest SE_e . However, as none of the statistical methods cannot give a full answer to the question which of the models is the best, it is the researcher who should choose the model on the basis of the results.

EXAMPLE (17.1) c.d. (*publisher.pqs file*)

To predict the gross profit from book sales a publisher wants to consider such variables as: production cost, advertising costs, direct promotion cost, the sum of discounts made, and the author's popularity. However, not all of those variables need to have a significant effect on profit. Let us try to select such a model of linear regression which will contain the optimum number of variables (from the perspective of statistics).

- **Manual model comparison.**

On the basis of the earlier constructed, full model we can suspect that the variables: direct promotion costs and the sum of discounts made have a small influence on the constructed model (i.e. those variables do not help predict the greatness of the profit). We will check if, from the perspective of statistics, the full model is better than the model from which the two variables have been removed.

Comparing Multiple Regression Models	
Analysis time	0.04sec.
Analysed variables	gross_profit;prod_c;advert_c
Significance level	0.05
Size	40
Number of variables in the model 1	5
Analysed variables	prod_c;advert_c;prom_c;reb
Standard error of estimation	8.086501
R	0.922483
R2	0.850974
Adjusted R2	0.829059
Number of variables in the model 2	3
Analysed variables	prod_c;advert_c;popular_au
Standard error of estimation	8.072538
R	0.918016
R2	0.842753
Adjusted R2	0.829649
F - comparing models	0.937892
DF1	2
DF2	34
p-value	0.401345

Model 1									
	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. er	
intercept	4.175186	4.772779	-5.524268	13.874639	0.874791	0.387825			
prod_c	2.560709	0.501507	1.541525	3.579894	5.106033	0.000013	0.422818	0.082807	
advert_c	1.998235	0.359065	1.268527	2.727943	5.565106	0.000003	0.461287	0.082889	
prom_c	4.668238	4.791644	-5.069555	14.40603	0.974245	0.336816	0.066242	0.067993	
rebates	1.423171	1.404811	-1.431749	4.278091	1.013069	0.318183	0.067478	0.066608	
popular_au	10.153717	2.782567	4.498861	15.808574	3.649047	0.000874	0.261561	0.071679	

Model 2								
	b coeff.	b error	-95% CI	+95% CI	t stat.	p-value	b stand.	b stand. er
intercept	8.85096	3.277587	2.203704	15.498215	2.70045	0.010486		
prod_c	2.519802	0.492781	1.520396	3.519209	5.113431	0.000011	0.416063	0.081367
advert_c	2.074037	0.350655	1.362876	2.785198	5.914752	0.000001	0.478786	0.080948
popular_at	9.921455	2.771625	4.30034	15.54257	3.579653	0.001007	0.255578	0.071397

It turns out that there is no basis for thinking that the full model is better than the reduced model (the value p of F-test which is used for comparing models is $p = 0.401345$). Additionally, the reduced model is slightly more adequate than the full model (for the reduced model $R_{adj}^2 = 0.82964880$, for the full model $R_{adj}^2 = 0.82905898$).

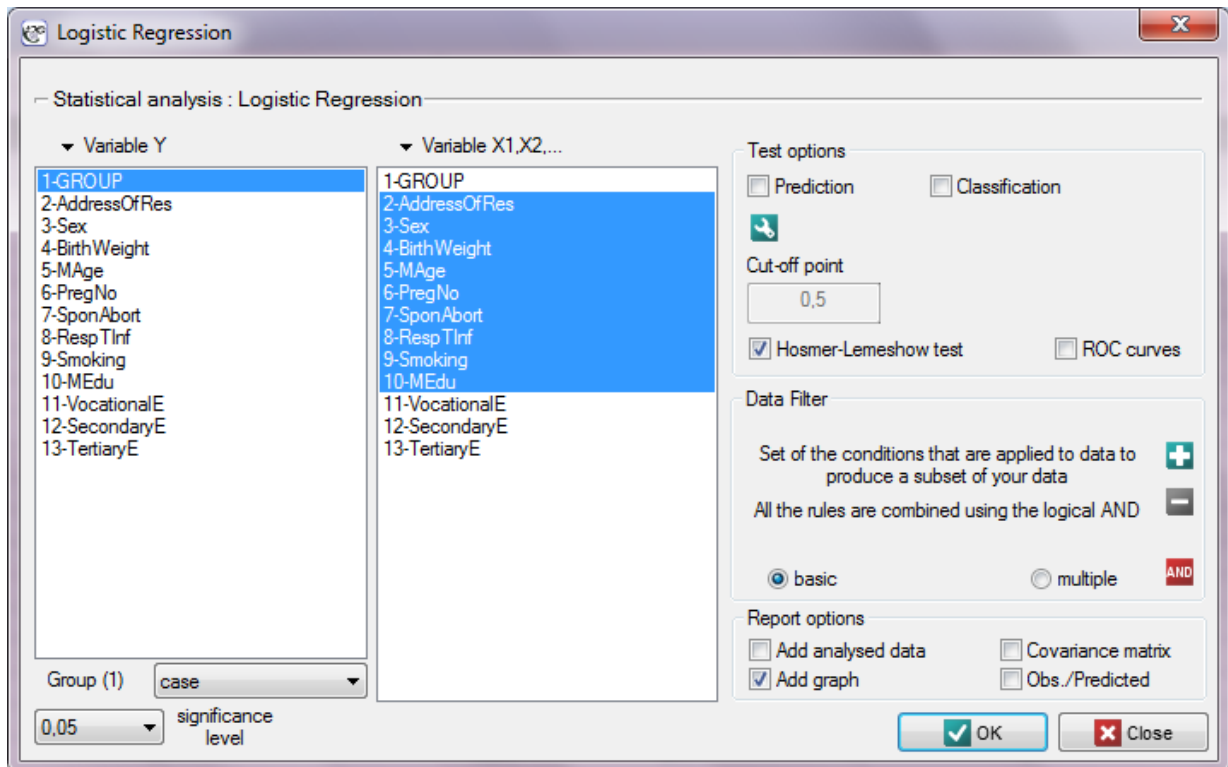
- **Automatic** model comparison.

In the case of automatic model comparison we receive very similar results. The best model is the one with the greatest coefficient R_{adj}^2 and the smallest standard estimation error SE_e . The best model we suggest is the model containing only 3 independent variables: the production cost, advertising costs, and the author's popularity.

On the basis of the analyses above, from the perspective of statistics, the optimum model is the model with the 3 most important independent variables: the production cost, advertising costs, and the author's popularity. However, the final decision which model to choose should be made by a person with specialist knowledge about the studied topic – in this case, the publisher. It ought to be remembered that the selected model should be constructed anew and its assumptions verified in the window Multiple regression.

17.4 LOGISTIC REGRESSION

The window with settings for Logistic Regression is accessed via the menu Statistics→Multidimensional Models→Logistic Regression



The constructed model of logistic regression (similarly to the case of multiple linear regression) allows the study of the effect of many independent variables (X_1, X_2, \dots, X_k) on one dependent variable (Y). This time, however, the dependent variable only assumes two values, e.g. ill/healthy, insolvent/solvent etc.

The two values are coded as (1)/(0), where:

- (1) –the distinguished value –possessing the feature
- (0) –not possessing the feature.

The function on which the model of logistic regression is based does not calculate the 2-level variable Y but the probability of that variable assuming the distinguished value:

$$P(Y = 1|X_1, X_2, \dots, X_k) = \frac{e^Z}{1 + e^Z}$$

where:

$P(Y = 1|X_1, X_2, \dots, X_k)$ –the probability of assuming the distinguished value (1) on condition that specific values of independent variables are achieved, the so-called probability predicted for 1.

Z is most often expressed in the form of a linear relationship:

$$Z = \beta_0 + \sum_{i=1}^k \beta_i X_i,$$

X_1, X_2, \dots, X_k –independent variables, explanatory,

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ –parameters.

Note!

Function Z can also be described with the use of a higher order relationship, e.g. a square relationship - in such a case we introduce into the model a variable containing the square of the independent variable X_i^2 .

Note!

Function Z can contain variable interactions - in such a case we introduce into the model a variable which is the result of multiplying the variables participating in the interaction, e.g. $X_1 \times X_2$.

The logit is the transformation of that model into the form:

$$\ln \left(\frac{P}{1-P} \right) = Z.$$

The matrices involved in the equation, for a sample of size n , are recorded in the following manner:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

In such a case, the solution of the equation is the vector of the estimates of parameters $\beta_0, \beta_1, \dots, \beta_k$ called **regression coefficients**:

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}.$$

The coefficients are estimated with the use of the **maximum likelihood method**, that is through the search for the maximum value of likelihood function L (in the program the Newton-Raphson iterative algorithm was used). On the basis of those values we can infer the magnitude of the effect of the independent variable (for which the coefficient was estimated) on the dependent variable.

There is a certain error of estimation for each coefficient. The magnitude of that error is estimated from the following formula:

$$SE_b = \sqrt{\text{diag}(H^{-1})_b},$$

where:

$\text{diag}(H^{-1})$ is the main diagonal of the covariance matrix.

Note!

When building the model you need remember that the number of observations should be ten times greater than or equal to the number of the estimated parameters of the model ($n \geq 10(k+1)$).

Note!

When building the model you need remember that the independent variables should not be multicollinear. In a case of multicollinearity estimation can be uncertain and the obtained error values very high. The multicollinear variables should be removed from the model or one independent variable should be built of them, e.g. instead of the multicollinear variables of mother age and father age one

can build the parents age variable.

Note!

The criterion of convergence of the function of the Newton-Raphson iterative algorithm can be controlled with the help of two parameters: the limit of convergence iteration (it gives the maximum number of iterations in which the algorithm should reach convergence) and the convergence criterion (it gives the value below which the received improvement of estimation shall be considered to be insignificant and the algorithm will stop).

17.4.1 Odds Ratio

Individual Odds Ratio

On the basis of many coefficients, for each independent variable in the model an easily interpreted measure is estimated, i.e. the individual Odds Ratio:

$$OR_i = e^{\beta_i}.$$

The received Odds Ratio expresses the change of the odds for the occurrence of the distinguished value (1) when the independent variable grows by 1 unit. The result is corrected with the remaining independent variables in the model so that it is assumed that they remain at a stable level while the studied variable is growing by 1 unit.

The OR value is interpreted as follows:

- $OR > 1$ means the stimulating influence of the studied independent variable on obtaining the distinguished value (1), i.e. it gives information about how much greater are the odds of the occurrence of the distinguished value (1) when the independent variable grows by 1 unit.
- $OR < 1$ means the destimulating influence of the studied independent variable on obtaining the distinguished value (1), i.e. it gives information about how much lower are the odds of the occurrence of the distinguished value (1) when the independent variable grows by 1 unit.
- $OR \approx 1$ means that the studied independent variable has no influence on obtaining the distinguished value (1).

Odds Ratio - the general formula

The PQStat program calculates the individual Odds Ratio. Its modification on the basis of a general formula makes it possible to change the interpretation of the obtained result.

The Odds Ratio for the occurrence of the distinguished state in a general case is calculated as the ratio of two odds. Therefore for the independent variable X_1 for Z expressed with a linear relationship we calculate:

the odds for the first category:

$$Odds(1) = \frac{P(1)}{1 - P(1)} = e^Z(1) = e^{\beta_0 + \beta_1 X_1(1) + \beta_2 X_2 + \dots + \beta_k X_k},$$

the odds for the second category:

$$Odds(2) = \frac{P(2)}{1 - P(2)} = e^Z(2) = e^{\beta_0 + \beta_1 X_1(2) + \beta_2 X_2 + \dots + \beta_k X_k}.$$

The Odds Ratio for variable X_1 is then expressed with the formula:

$$\begin{aligned}
 OR_1(2)/(1) &= \frac{Odds(2)}{Odds(1)} = \frac{e^{\beta_0 + \beta_1 X_1(2) + \beta_2 X_2 + \dots + \beta_k X_k}}{e^{\beta_0 + \beta_1 X_1(1) + \beta_2 X_2 + \dots + \beta_k X_k}} \\
 &= e^{\beta_0 + \beta_1 X_1(2) + \beta_2 X_2 + \dots + \beta_k X_k - [\beta_0 + \beta_1 X_1(1) + \beta_2 X_2 + \dots + \beta_k X_k]} \\
 &= e^{\beta_1 X_1(2) - \beta_1 X_1(1)} = e^{\beta_1 [X_1(2) - X_1(1)]} = \\
 &= (e^{\beta_1})^{[X_1(2) - X_1(1)]}.
 \end{aligned}$$

Example

If the independent variable is age expressed in years, then the difference between neighboring age categories such as 25 and 26 years is 1 year ($X_1(2) - X_1(1) = 26 - 25 = 1$). In such a case we will obtain the individual Odds Ratio:

$$OR = (e^{\beta_1})^1,$$

which expresses the degree of change of the odds for the occurrence of the distinguished value if the age is changed by 1 year.

The odds ratio calculated for non-neighboring variable categories, such as 25 and 30 years, will be a five-year Odds Ratio, because the difference $X_1(2) - X_1(1) = 30 - 25 = 5$. In such a case we will obtain the five-year Odds Ratio:

$$OR = (e^{\beta_1})^5,$$

which expresses the degree of change of the odds for the occurrence of the distinguished value if the age is changed by 5 years.

Note!

If the analysis is made for a non-linear model or if interaction is taken into account, then, on the basis of a general formula, we can calculate an appropriate Odds Ratio by changing the formula which expresses Z .

17.4.2 Model verification

Statistical significance of particular variables in the model (significance of the Odds Ratio)

On the basis of the coefficient and its error of estimation we can infer if the independent variable for which the coefficient was estimated has a significant effect on the dependent variable. For that purpose we use Wald test.

Hypotheses:

$$\begin{aligned}
 \mathcal{H}_0 : \beta_i &= 0, & \text{or, equivalently: } \mathcal{H}_0 : OR_i &= 1, \\
 \mathcal{H}_1 : \beta_i &\neq 0. & \mathcal{H}_1 : OR_i &\neq 1.
 \end{aligned}$$

The Wald test statistics is calculated according to the formula:

$$\chi^2 = \left(\frac{b_i}{SE_{b_i}} \right)^2$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with 1 degree of freedom. On the basis of test statistics, p value is estimated and then compared with the significance level α :

if $p \leq \alpha \implies$ we reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

The quality of the constructed model of multiple linear regression can be evaluated with the help of several measures

- **Pseudo R^2** – is a goodness of fit measure of the model (an equivalent of the coefficient of multiple determination R^2 defined for multiple linear regression). The value of that coefficient falls within the range of $< 0; 1$), where values close to 1 mean excellent goodness of fit of a model, 0 – a complete lack of fit. Coefficient R_{Pseudo}^2 is calculated according to the formula:

$$R_{Pseudo}^2 = 1 - \frac{\ln L_{FM}}{\ln L_0}.$$

where:

L_{FM} – the maximum value of likelihood function of a full model (with all variables),
 L_0 – the maximum value of likelihood function of a model which only contains a intercept.

As coefficient R_{Pseudo}^2 never assumes value 1 and is sensitive to the amount of variables in the model, its corrected value is calculated:

$$R_{Nagelkerke}^2 = \frac{1 - e^{-(2/n)(\ln L_{FM} - \ln L_0)}}{1 - e^{(2/n) \ln L_0}} \quad \text{lub} \quad R_{Cox-Snell}^2 = 1 - e^{\frac{(-2 \ln L_0) - (-2 \ln L_{FM})}{n}}.$$

- **Statistical significance of all variables in the model**

The basic tool for the evaluation of the significance of all variables in the model is **the Likelihood Ratio test**. The test verifies the hypothesis:

$$\begin{aligned} \mathcal{H}_0 : & \quad \text{all } \beta_i = 0, \\ \mathcal{H}_1 : & \quad \text{there is } \beta_i \neq 0. \end{aligned}$$

The test statistic has the form presented below:

$$\chi^2 = -2 \ln(L_0/L_{FM}) = -2 \ln(L_0) - (-2 \ln(L_{FM})).$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with k degrees of freedom.

On the basis of **test statistics, p value** is estimated and then compared with α :

if $p \leq \alpha \implies$ we reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

- **Hosmer-Lemeshow test** – The test compares, for various subgroups of data, the observed rates of occurrence of the distinguished value O_g and the predicted probability E_g . If O_g and E_g are close enough then one can assume that an adequate model has been built.

For the calculation the observations are first divided into G subgroups – usually deciles ($G = 10$).

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : & \quad O_g = E_g \text{ for all categories,} \\ \mathcal{H}_1 : & \quad O_g \neq E_g \text{ for at least one category.} \end{aligned}$$

The test statistic has the form presented below:

$$H = \sum_{g=1}^G \frac{(O_g - E_g)^2}{E_g(1 - \frac{E_g}{N_g})},$$

where:

N_g –the number of observations in group g .

The statistic asymptotically (for large sizes) has the χ^2 distribution with $G - 2$ degrees of freedom.

On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

- **AUC - the area under the ROC curve** –The ROC curve built on the basis of the value of the dependent variable, and the predicted probability of dependent variable P , allows to evaluate the ability of the constructed logistic regression model to classify the cases into two groups: (1) and (0). The constructed curve, especially the area under the curve, presents the classification quality of the model. When the ROC curve overlaps with the diagonal $y = x$, then the decision about classifying a case within a given class (1) or (0), made on the basis of the model, is as good as a random division of the studied cases into the groups. The classification quality of a model is good when the curve is much above the diagonal $y = x$, that is when the area under the ROC curve is much larger than the area under the $y = x$ line, i.e. it is greater than 0.5

Hypotheses:

$$\begin{aligned} \mathcal{H}_0 : \quad AUC &= 0.5, \\ \mathcal{H}_1 : \quad AUC &\neq 0.5. \end{aligned}$$

The test statistic has the form presented below:

$$Z = \frac{AUC - 0.5}{SE_{0.5}},$$

where:

$SE_{0.5}$ –area error.

Statistics Z asymptotically (for large sizes) has the normal distribution.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

Additionally, for ROC curve the suggested value of the **cut-off point** of the predicted probability is given, together with the table of sensitivity and specificity for each possible cut-off point.

Note!

More possibilities of calculating a cut-off point are offered by module **ROC curve**. The analysis is made on the basis of observed values and predicted probability obtained in the analysis of Logistic Regression.

- **Classification**

On the basis of the selected cut-off point of predicted probability we can change the classification quality. By default the cut-off point has the value of 0.5. The user can change the value into any value from the range of (0.1), e.g. the value suggested by the ROC curve.

As a result we shall obtain the classification table and the percentage of properly classified cases, the percentage of properly classified (0) –specificity, and the percentage of properly classified (1) –sensitivity.

Prediction on the basis of the model

On the basis of a selected cut-off point of predicted probability and of the given values of independent variables we can calculate the predicted value of the dependent value (0) or (1). By default the cut-off point has the value of 0.5. The user can change the value into any value from the range of (0.1), e.g. the value suggested by the ROC curve.

EXAMPLE 17.2. (anomaly.pqs file)

Studies have been conducted for the purpose of identifying the risk factors for a certain rare congenital anomaly in children. 395 mothers of children with that anomaly and 375 of healthy children have participated in that study. The gathered data are: address of residence, child's sex, child's weight at birth, mother's age, number of pregnancy, previous spontaneous abortions, respiratory tract infections, smoking, mother's education.

We construct a logistic regression model to check which variables may have a significant influence on the occurrence of the anomaly. The dependent variable is the column GROUP, the distinguished values in that variable as 1 are the "cases", that are mothers of children with anomaly. The following 9 variables are independent variables:

AddressOfRes (2=city/1=village),
 Sex (1=male/0=female),
 BirthWeight (in kilograms, with an accuracy of 0.5 kg),
 MAge (in years),
 PregNo (which pregnancy is the child from),
 SponAbort (1=yes/0=no),
 RespTInf (1=yes/0=no),
 Smoking (1=yes/0=no),
 MEdu (1=primary or lower/2=vocational/3=secondary/4=tertiary).

Logistic Regression	
Analysis time	0.65sec.
Analysed variables	AddressOfRes;Sex;BirthWeig
Count of missing data	89
Significance level	0.05
Convergence has been reached	
Size	678
Number of estimated parameters	10
Frequency 0 (control)	346
Frequency 1 (case)	332
Likelihood ratio test	
Log Likelihood	-418.000866
-2 Log Likelihood	836.001732
Log Likelihood (intercept)	-469.809235
-2 Log Likelihood (intercept)	939.618471
Chi-square statistic	103.616739
Degrees of freedom	9
p-value	<0.000001
Pseudo R2	0.110275
R2(Nagelkerke)	0.188989
R2(Coxa-Snell)	0.141722
Hosmer-Lemeshow test	
Chi-square statistic	9.855715
Degrees of freedom	8
p-value	0.275299

Model	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	odds ratio	-95% CI	+95% CI
intercept	1.473902	0.664907	0.170709	2.777096	4.913783	0.026643	4.366241	1.186146	16.072277
AddressOfRes	-0.040877	0.171507	-0.377024	0.29527	0.056807	0.811616	0.959947	0.685899	1.343489
Sex	0.464687	0.170064	0.131368	0.798005	7.46616	0.006287	1.591515	1.140387	2.221106
BirthWeight	-0.307868	0.131076	-0.564772	-0.050963	5.516717	0.018836	0.735013	0.56849	0.950314
MAge	-0.033758	0.018671	-0.070353	0.002837	3.268947	0.070603	0.966805	0.932064	1.002841
PregNo	0.293138	0.100444	0.096271	0.490005	8.517193	0.003518	1.340628	1.101058	1.632324
SponAbort	-0.433693	0.303193	-1.02794	0.160554	2.046102	0.152596	0.648111	0.357743	1.174161
RespTInf	1.495785	0.277773	0.95136	2.040209	28.99738	<0.000001	4.462837	2.589229	7.69222
Smoking	1.490982	0.411868	0.683736	2.298227	13.104768	0.000295	4.441453	1.981266	9.956516
MEdu	-0.183437	0.101185	-0.381755	0.014881	3.286588	0.069848	0.832404	0.682662	1.014992

The quality of model goodness of fit is not high ($R^2_{pseudo} = 0.11$, $R^2_{Nagelkerke} = 0.19$ i $R^2_{Cox-Snell} = 0.14$). At the same time the model is statistically significant (value $p < 0.000001$ of the Likelihood Ratio test), which means that a part of the independent variables in the model is statistically significant. The result of the Hosmer-Lemeshow test points to a lack of significance ($p = 0.2753$). However, in the case of the Hosmer-Lemeshow test we ought to remember that a lack of significance is desired as it indicates a similarity of the observed sizes and of predicted probability.

An interpretation of particular variables in the model starts from checking their significance. In this case the variables which are significantly related to the occurrence of the anomaly are:

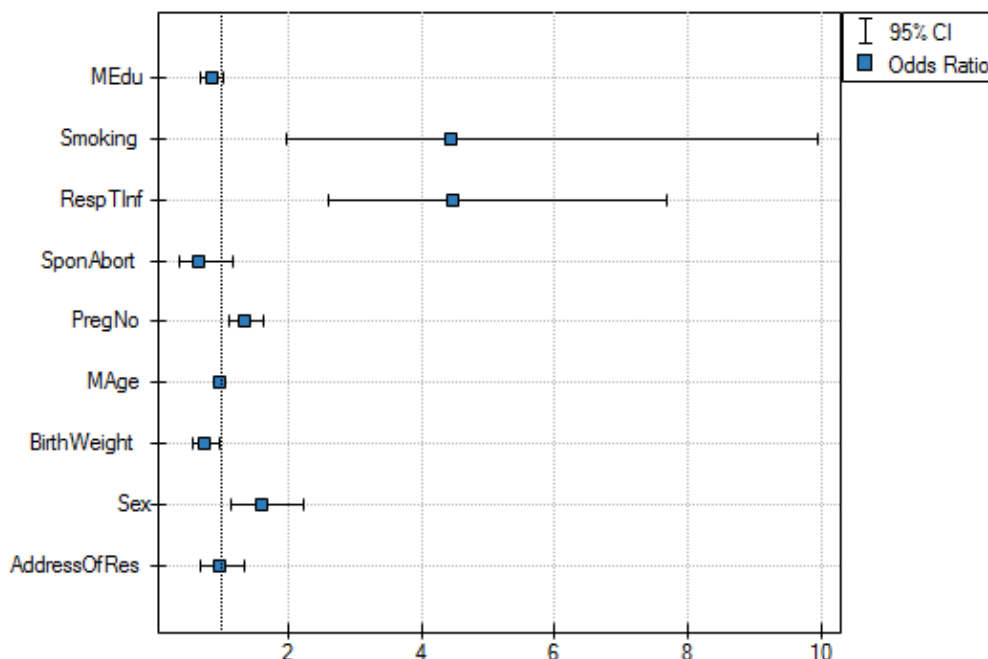
Sex: $p = 0.0063$,
 BirthWeight: $p = 0.0188$,
 PregNo: $p = 0.0035$,
 RespTInf: $p < 0.000001$,
 Smoking: $p = 0.0003$.

The studied congenital anomaly is a rare anomaly but the odds of its occurrence depend on the variables listed above in the manner described by the odds ratio:

- variable Sex: $OR[95\%CI] = 1.60[1.14; 2.22]$ –the odds of the occurrence of the anomaly in a boy is 1.6 times greater than in a girl;
- variable BirthWeight: $OR[95\%CI] = 0.74[0.57; 0.95]$ –the higher the birth weight the smaller the odds of the occurrence of the anomaly in a child;
- variable PregNo: $OR[95\%CI] = 1.34[1.10; 1.63]$ –the odds of the occurrence of the anomaly in a child is 1.34 times greater with each subsequent pregnancy;
- variable RespTInf: $OR[95\%CI] = 4.46[2.59; 7.69]$ –the odds of the occurrence of the anomaly in a child if the mother had a respiratory tract infection during the pregnancy is 4.46 times greater than in a mother who did not have such an infection during the pregnancy;
- variable Smoking: $OR[95\%CI] = 4.44[1.98; 9.96]$ –a mother who smokes when pregnant increases the risk of the occurrence of the anomaly in her child 4.44 times.

In the case of statistically insignificant variables the confidence interval for the Odds Ratio contains 1 which means that the variables neither increase nor decrease the odds of the occurrence of the studied anomaly. Therefore, we cannot interpret the obtained ratio in a manner similar to the case of statistically significant variables.

The influence of particular independent variables on the occurrence of the anomaly can also be described with the help of a chart concerning the odds ratio:



Note!

An independent variable with a few categories can be considered in the model as dummy variables. In such a case, before the commencement of the analysis one should divide that variable into a few dummy variables with 2 categories.

EXAMPLE 17.2 c.d. (anomaly.pqs)

Let us once more construct a logistic regression model, however, this time let us divide the variable mother's education into dummy variables. With this operation we lose the information about the ordering of the category of education but we gain the possibility of a more in-depth analysis of particular categories. The division into dummy variables was made by creating 3 variables concerning mother's education:

VocationalE (1=yes/0=no),
 SecondaryE (1=yes/0=no),
 TertiaryE (1=yes/0=no).

The primary education variable is missing as it will constitute the reference category.

Model									
	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	odds ratio	-95% CI	+95% CI
intercept	1.665115	0.693346	0.306183	3.024048	5.767521	0.016325	5.286283	1.358231	20.574401
AddressOfRes	-0.046576	0.172997	-0.385643	0.292491	0.072484	0.787754	0.954492	0.680014	1.339761
Sex	0.438115	0.171101	0.102942	0.773288	6.563484	0.010409	1.549783	1.108427	2.166879
BirthWeight	-0.295937	0.13156	-0.55379	-0.038084	5.059981	0.024485	0.743835	0.574768	0.962633
MAge	-0.034094	0.018834	-0.071008	0.00282	3.277034	0.070256	0.96648	0.931454	1.002824
PregNo	0.299655	0.101019	0.101662	0.497648	8.799159	0.003014	1.349394	1.10701	1.644848
SponAbort	-0.491751	0.306896	-1.093256	0.109755	2.567479	0.109081	0.611555	0.335124	1.116004
RespTInf	1.487811	0.27766	0.943608	2.032014	28.712438	<0.000001	4.427393	2.569235	7.629434
Smoking	1.457938	0.414507	0.64552	2.270356	12.371291	0.000436	4.29709	1.906978	9.682849
VocationalE	-0.682289	0.344821	-1.358125	-0.006452	3.915161	0.047852	0.505459	0.257142	0.993568
SecondaryE	-0.871537	0.332673	-1.523565	-0.21951	6.863349	0.008798	0.418308	0.217934	0.802912
TertiaryE	-0.79081	0.359937	-1.496274	-0.085347	4.827158	0.028015	0.453477	0.223963	0.918194

Logistic Regression	
Analysis time	0.28sec.
Analysed variables	AddressOfRes;Sex;BirthWeig
Count of missing data	89
Significance level	0.05
Convergence has been reached	
Size	678
Number of estimated parameters	12
Frequency 0 (control)	346
Frequency 1 (case)	332
Likelihood ratio test	
Log Likelihood	-416.06069
-2 Log Likelihood	832.12138
Log Likelihood (intercept)	-469.809235
-2 Log Likelihood (intercept)	939.618471
Chi-square statistic	107.49709
Degrees of freedom	11
p-value	<0.000001
Pseudo R2	0.114405
R2(Nagelkerke)	0.195521
R2(Coxa-Snella)	0.14662
Hosmer-Lemeshow test	
Chi-square statistic	6.720908
Degrees of freedom	8
p-value	0.567022

As a result the variables which describe education become statistically significant. The goodness of fit of the model does not change much but the manner of interpretation of the the odds ratio for education does change:

Variable	OR[95%CI]
Primary education	reference category
Vocational education	0.51[0.26; 0.99]
Secondary education	0.42[0.22; 0.80]
Tertiary education	0.45[0.22; 0.92]

The odds of the occurrence of the studied anomaly in each education category is always compared with the odds of the occurrence of the anomaly in the case of primary education. We can see that for more educated the mother, the odds is lower. For a mother with:

- vocational education the odds of the occurrence of the anomaly in a child is 0.51 of the odds for a mother with primary education;
- secondary education the odds of the occurrence of the anomaly in a child is 0.42 of the odds for a mother with primary education;
- tertiary education the odds of the occurrence of the anomaly in a child is 0.45 of the odds for a mother with primary education;

EXAMPLE 17.3. (task.pqs file)

An experiment has been made with the purpose of studying the ability to concentrate of a group of

adults in an uncomfortable situation. 130 people have taken part in the experiment. Each person was assigned a certain task the completion of which required concentration. During the experiment some people were subject to a disturbing agent in the form of temperature increase to 32 degrees Celsius. The participants were also asked about their address of residence, sex, age, and education. The time for the completion of the task was limited to 45 minutes. In the case of participants who completed the task before the deadline, the actual time devoted to the completion of the task was recorded.

Variable SOLUTION (yes/no) contains the result of the experiment, i.e. the information about whether the task was solved correctly or not. The remaining variables which could have influenced the result of the experiment are:

ADDRESSOFRES (1=city/0=village),
 SEX (1=female/0=male),
 AGE (in years),
 EDUCATION (1=primary, 2=vocational, 3=secondary, 4=tertiary),
 TIME needed for the completion of the task (in minutes),
 DISTURBANCES (1=yes/0=no).

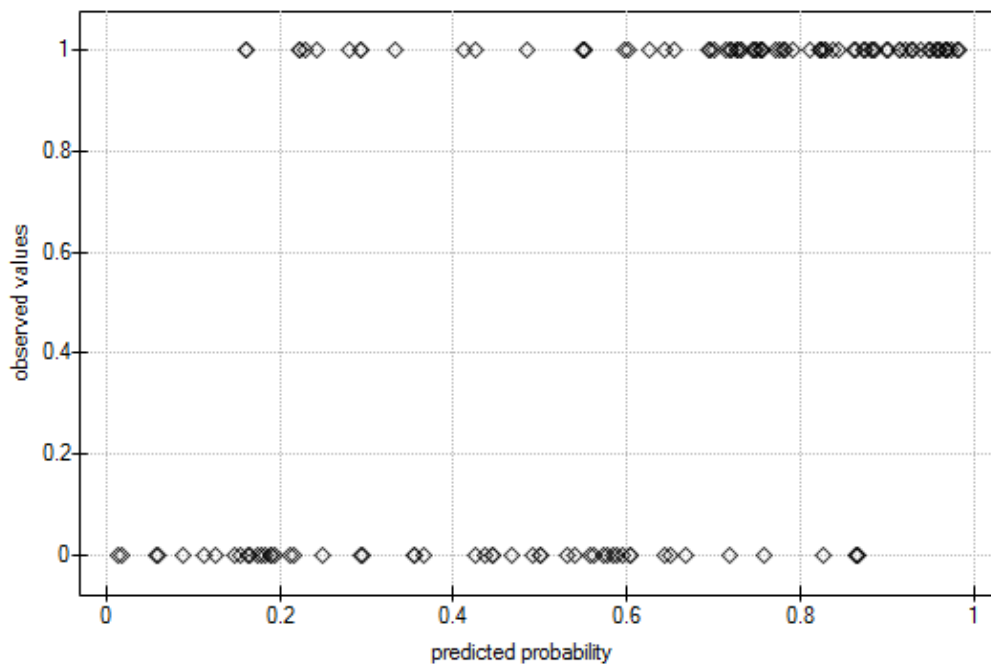
On the basis of all those variables a logistic regression model was built in which the distinguished state of the variable SOLUTION was set to "yes".

Logistic Regression	
Analysis time	0.20sec.
Analysed variables	ADDRESSOFRES;SEX;AGE;E
Significance level	0.05
Convergence has been reached	
Size	130
Number of estimated parameters	7
Frequency 0 (no)	53
Frequency 1 (yes)	77
Likelihood ratio test	
Log Likelihood	-64.354117
-2 Log Likelihood	128.708234
Log Likelihood (intercept)	-87.88099
-2 Log Likelihood (intercept)	175.761979
Chi-square statistic	47.053745
Degrees of freedom	6
p-value	<0.000001
Pseudo R2	0.267713
R2(Nagelkerke)	0.409674
R2(Coxa-Snella)	0.303684
Hosmer-Lemeshow test	
Chi-square statistic	11.548615
Degrees of freedom	8
p-value	0.172508

The adequacy quality is described by the coefficients: $R_{Pseudo}^2 = 0.27$, $R_{Nagelkerke}^2 = 0.41$ i $R_{Cox-Snell}^2 = 0.30$. The sufficient adequacy is also indicated by the result of the Hosmer-Lemeshow test ($p = 0.1725$). The whole model is statistically significant, which is indicated by the result of the Likelihood Ratio test ($p < 0.000001$).

Model									
	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	odds ratio	-95% CI	+95% CI
intercept	7.230601	1.870134	3.565206	10.895997	14.948697	0.00011	1381.0525	35.346728	53959.905
ADDRESSOFRES	-0.453242	0.450524	-1.336253	0.429769	1.012102	0.3144	0.635564	0.262829	1.536902
SEX	-0.454788	0.451304	-1.339327	0.429751	1.015501	0.313589	0.634582	0.262022	1.536875
AGE	-0.100896	0.03159	-0.162812	-0.03898	10.200921	0.001404	0.904027	0.849751	0.96177
EDUCATION	0.455928	0.241805	-0.018	0.929857	3.555199	0.059359	1.577637	0.982161	2.534146
TIME	-0.089395	0.027609	-0.143507	-0.035282	10.483921	0.001204	0.914484	0.866314	0.965333
DISTURBANCES	-1.924	0.475056	-2.855092	-0.992908	16.402912	0.000051	0.146022	0.057551	0.370498

The observed values and predicted probability can be observed on the chart:



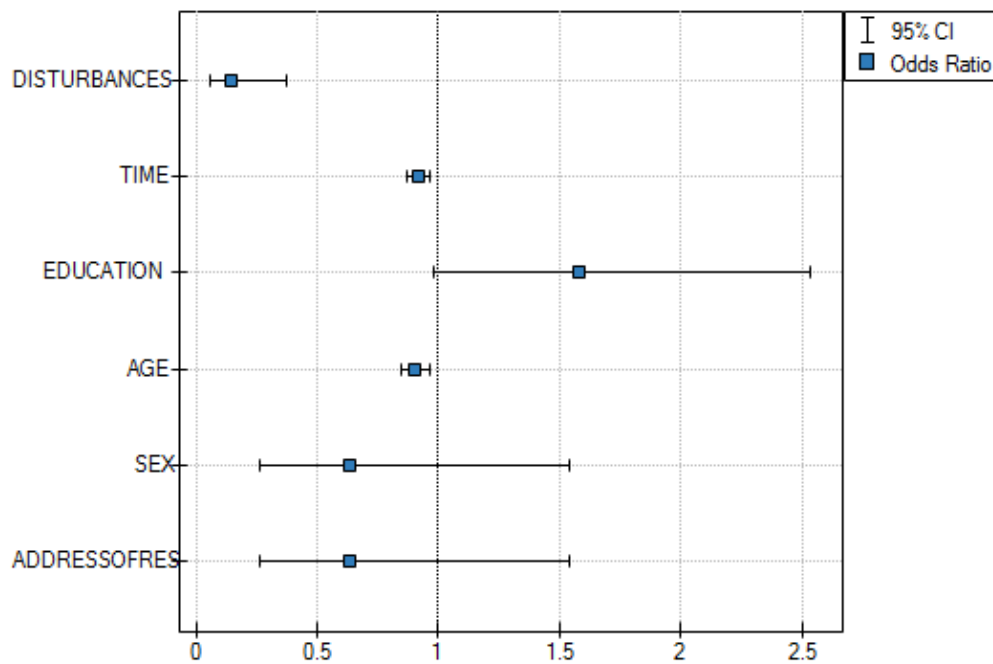
In the model the variables which have a significant influence on the result are:

AGE: $p = 0.0014$,
 TIME: $p = 0.0012$,
 DISTURBANCES: $p = 0.0001$.

What is more, the younger the person who solves the task the shorter the time needed for the completion of the task, and if there is no disturbing agent, the probability of correct solution is greater:

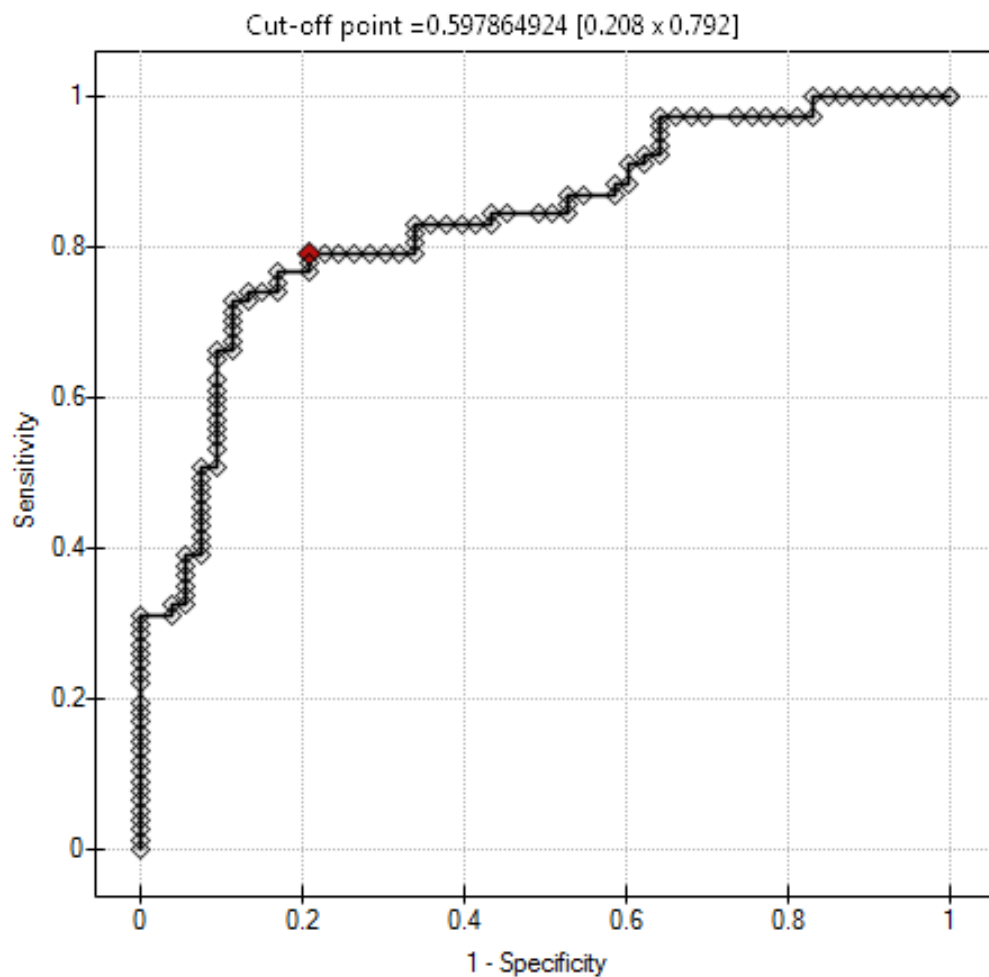
AGE: $OR[95\%CI] = 0.90[0.85; 0.96]$,
 TIME: $OR[95\%CI] = 0.91[0.87; 0.97]$,
 DISTURBANCES: $OR[95\%CI] = 0.15[0.06; 0.37]$.

The obtained results of the Odds Ratio are presented on the chart below:



Should the model be used for prediction, one should pay attention to the quality of classification. For that purpose we calculate the ROC curves.

ROC curves (DeLong's method)	
AUC	0.834599
SE(AUC)	0.035432
-95% CI	0.765153
+95% CI	0.904045
Z statistic	6.469391
p-value	<0.000001
Cut-off point	0.597865



The result seems satisfactory. The area under the curve is $AUC = 0.83$ and is statistically greater than 0.5 ($p < 0.000001$), so classification is possible on the basis of the constructed model. The suggested cut-off point for the ROC curve is 0.60 and is slightly higher than the standard level used in regression, i.e. 0.5. Classification made on the basis of that cut-off point yields 78.46% correctly classified cases, of which the correctly classified "yes" values constitute 77.92% (sensitivity[95%CI] = 77.92%[67.02%; 86.58%]), the "no" values constitute 79.25% (specificity[95%CI] = 79.25%[65.89%; 89.16%]).

Classification		Observed value	
Predicted value		1	0
	1	60	11
	0	17	42
Cut-off point	0.6		
% correct	78.46%		
Sensitivity (% cc	77.92%		
-95% CI	67.02%		
+95% CI	86.58%		
Specificity (% cc	79.25%		
-95% CI	65.89%		
+95% CI	89.16%		

We can finish the analysis of classification at this stage or, if the result is not satisfactory, we can make a more detailed analysis of the ROC curve in module [ROC curve](#).

As we have assumed that classification on the basis of that model is satisfactory we can calculate the predicted value of a dependent variable for any conditions. Let us check what odds of solving the task has a person whose:

ADDRESSOFRES (1=city),
SEX (1=female),
AGE (50 years),
EDUCATION (1=primary),
TIME needed for the completion of the task (20 minutes),
DISTURBANCES (1=yes).

For that purpose, on the basis of the value of coefficient b , we calculate the predicted probability (probability of receiving the answer "yes" on condition of defining the values of dependent variables):

$$\begin{aligned}
 P(Y = yes | ADDRESSOFRES, SEX, AGE, EDUCATION, TIME, DISTURBANCES) &= \\
 &= \frac{e^{7.23 - 0.45ADDRESSOFRES - 0.45SEX - 0.1AGE + 0.46EDUCATION - 0.09TIME - 1.92DISTURBANCES}}{1 + e^{7.23 - 0.45ADDRESSOFRES - 0.45SEX - 0.1AGE + 0.46EDUCATION - 0.09TIME - 1.92DISTURBANCES}} = \\
 &= \frac{e^{7.231 - 0.453 \cdot 1 - 0.455 \cdot 1 - 0.101 \cdot 50 + 0.456 \cdot 1 - 0.089 \cdot 20 - 1.924 \cdot 1}}{1 + e^{7.231 - 0.453 \cdot 1 - 0.455 \cdot 1 - 0.101 \cdot 50 + 0.456 \cdot 1 - 0.089 \cdot 20 - 1.924 \cdot 1}}
 \end{aligned}$$

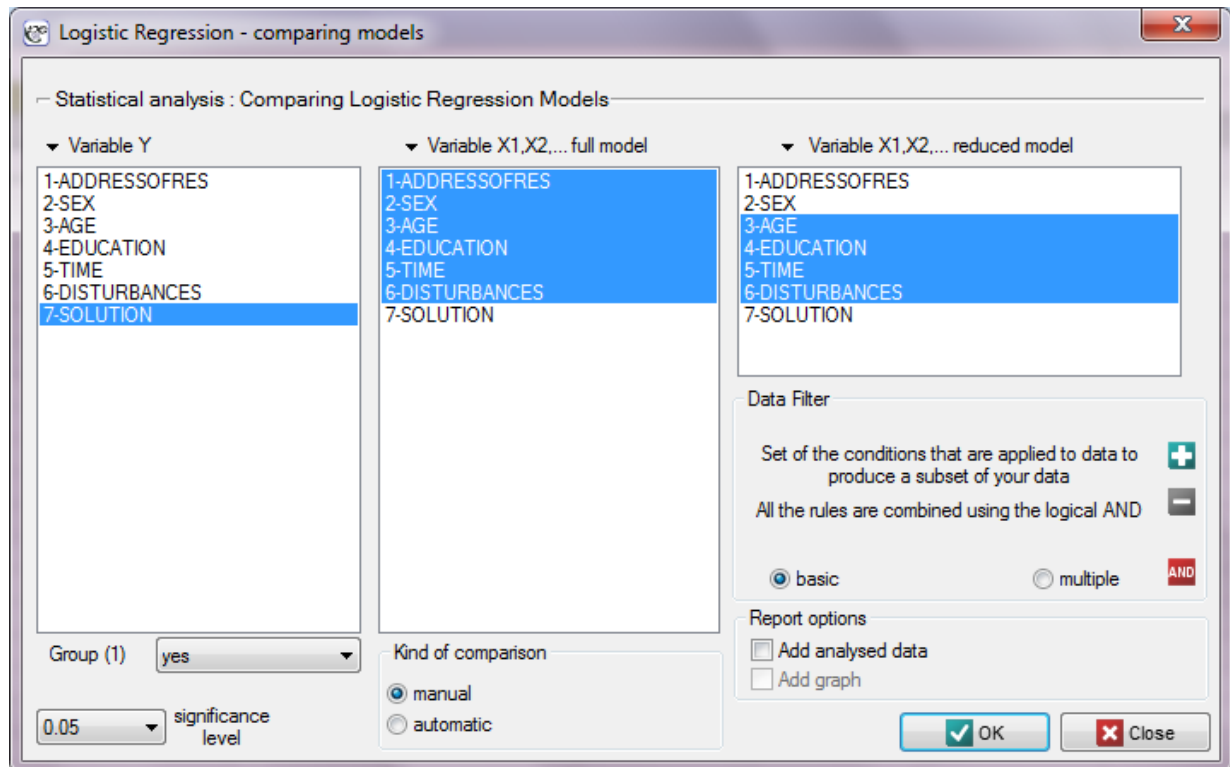
As a result of the calculation the program will return the result:

Prediction	
1-ADDRESSOFRES	1
2-SEX	1
3-AGE	50
4-EDUCATION	1
5-TIME	20
6-DISTURBANCE	1
Cut-off point	0.6
pred. prob.	0.121512
Pred. Y	0

The obtained probability of solving the task is equal to 0.1215, so, on the basis of the cut-off 0.60, the predicted result is 0 –which means the task was not solved correctly.

17.5 COMPARISON OF LOGISTIC REGRESSION MODELS

The window with settings for model comparison is accessed via the menu Statistics→Multidimensional models→Logistic regression – comparing models



Due to the possibility of simultaneous analysis of many independent variables in one logistic regression model, similarly to the case of multiple linear regression, there is a problem of selection of an optimum model. When choosing independent variables one has to remember to put into the model variables strongly correlated with the dependent variable and weakly correlated with one another.

When comparing models with various numbers of independent variables we pay attention to goodness of fit of the model (R_{Pseudo}^2 , $R_{Nagelkerke}^2$, $R_{Cox-Snell}^2$). For each model we also calculate the maximum of likelihood function which we later compare with the use of the Likelihood Ratio test.

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 &: L_{FM} = L_{RM}, \\ \mathcal{H}_1 &: L_{FM} \neq L_{RM},\end{aligned}$$

where:

L_{FM} , L_{RM} – the maximum of likelihood function in compared models (full and reduced).

The test statistic has the form presented below:

$$\chi^2 = -2 \ln(L_{RM}/L_{FM}) = -2 \ln(L_{RM}) - (-2 \ln(L_{FM}))$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with $df = k_{FM} - k_{RM}$ degrees of freedom, where k_{FM} i k_{RM} is the number of estimated parameters in compared models.

On the basis of test statistics, p value is estimated and then compared with α :

if $p \leq \alpha \implies$ we reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

We make the decision about which model to choose on the basis of the size R_{Pseudo}^2 , $R_{Nagelkerke}^2$, $R_{Cox-Snell}^2$ and the result of the Likelihood Ratio test which compares the subsequently created (neighboring) models. If the compared models do not differ significantly, we should select the one with a smaller number of variables. This is because a lack of a difference means that the variables present in the full model but absent in the reduced model do not carry significant information. However, if the difference is statistically significant, it means that one of them (the one with the greater number of variables, with a greater R^2) is significantly better than the other one.

In the program PQStat the comparison of models can be done manually or automatically.

- **Manual** model comparison – construction of 2 models:
 - a full model – a model with a greater number of variables,
 - a reduced model – a model with a smaller number of variables – such a model is created from the full model by removing those variables which are superfluous from the perspective of studying a given phenomenon.

The choice of independent variables in the compared models and, subsequently, the choice of a better model on the basis of the results of the comparison, is made by the researcher.

- **Automatic** model comparison is done in several steps:
 - step 1 Constructing the model with the use of all variables.
 - step 2 Removing one variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.
 - step 3 A comparison of the full and the reduced model.
 - step 4 Removing another variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.
 - step 5 A comparison of the previous and the newly reduced model.
 - ...

In that way numerous, ever smaller models are created. The last model only contains 1 independent variable.

EXAMPLE 17.3 c.d. (task.pqs file)

In the experiment made with the purpose to study the concentration abilities a logistic regression model was constructed on the basis of the following variables:

dependent variable: SOLUTION (yes/no) - information about whether the task was correctly solved or not;

independent variables:

ADDRESSOFRES (1=city/0=village),

SEX (1=female/0=male),

AGE (in years),

EDUCATION (1=primary, 2=vocational, 3=secondary, 4=tertiary),

TIME needed for the completion of the task (in minutes),

DISTURBANCES (1=yes/0=no).

Let us check if all independent variables are indispensable in the model.

- **Manual** model comparison.

On the basis of the previously constructed full model we can suspect that the variables: ADDRESSOFRES and SEX have little influence on the constructed model (i.e. we cannot successfully make classifications on the basis of those variables). Let us check if, from the statistical point of view, the full model is better than the model from which the two variables have been removed.

Comparing Logistic Regression Models	
Analysis time	0.15sec.
Analysed variables	ADDRESSOFRES;SEX;AGE;E
Significance level	0.05
Size	130
Number of variables in the model 1	7
Analysed variables	ADDRESSOFRES;SEX;AGE;E
-2 Log Likelihood	128.708234
Pseudo R2	0.267713
R2(Nagelkerke)	0.409674
R2(Coxa-Snella)	0.303684
Number of variables in the model 2	5
Analysed variables	AGE;EDUCATION;TIME;DIST
-2 Log Likelihood	131.082274
Pseudo R2	0.254206
R2(Nagelkerke)	0.392363
R2(Coxa-Snella)	0.290851
Chi-square - Comparing models	
Chi-square statistic	2.37404
Degrees of freedom	2
p-value	0.305129

Model 1										
	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	Odds Ratio	-95% CI	+95% CI	
intercept	7.230601	1.870134	3.565206	10.895997	14.948697	0.00011	1381.0525	35.346728	53959.905	
ADDRESSOFRES	-0.453242	0.450524	-1.336253	0.429769	1.012102	0.3144	0.635564	0.262829	1.536902	
SEX	-0.454788	0.451304	-1.339327	0.429751	1.015501	0.313589	0.634582	0.262022	1.536875	
AGE	-0.100896	0.03159	-0.162812	-0.03898	10.200921	0.001404	0.904027	0.849751	0.96177	
EDUCATION	0.455928	0.241805	-0.018	0.929857	3.555199	0.059359	1.577637	0.982161	2.534146	
TIME	-0.089395	0.027609	-0.143507	-0.035282	10.483921	0.001204	0.914484	0.866314	0.965333	
DISTURBA	-1.924	0.475056	-2.855092	-0.992908	16.402912	0.000051	0.146022	0.057551	0.370498	

Model 2										
	b coeff.	b error	-95% CI	+95% CI	Wald stat.	p-value	Odds Ratio	-95% CI	+95% CI	
intercept	6.782101	1.821888	3.211267	10.352936	13.857511	0.000197	881.91991	24.81049	31348.947	
AGE	-0.106345	0.031611	-0.168301	-0.044389	11.317897	0.000768	0.899114	0.845099	0.956582	
EDUCATION	0.50406	0.23731	0.038941	0.969178	4.511625	0.033665	1.655428	1.039709	2.635777	
TIME	-0.083768	0.026838	-0.13637	-0.031166	9.741995	0.001801	0.919644	0.872519	0.969315	
DISTURBA	-1.847732	0.46198	-2.753195	-0.942268	15.99675	0.000063	0.157594	0.063724	0.389743	

The results of the Likelihood Ratio test ($p = 0.3051$) indicates that there is no basis for believing that a full model is better than a reduced one. Therefore, with a slight worsening of model adequacy, the address of residence and the sex can be omitted.

Note!

The comparison of both models with respect to their ability to classify can be made by comparing ROC curves for those models. For that purpose we use the module Dependent ROC Curves - a comparison described in Chapter ??.

- **Automatic** model comparison.

In the case of automatic model comparison we receive very similar results. The best model is the one constructed on the basis of independent variables: AGE, EDUCATION, TIME needed for the completion of the task, DISTURBANCES.

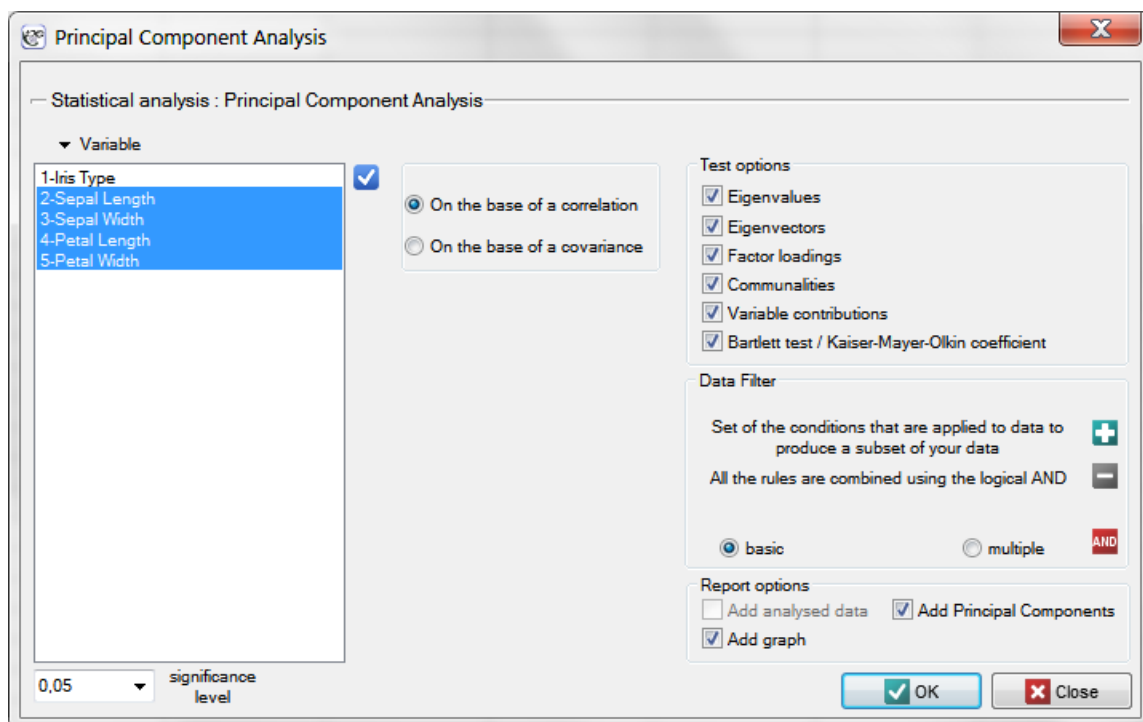
On the basis of the analyses above, from the statistical point of view, the optimumm model is a model with the 4 most important independent variables: AGE, EDUCATION, TIME needed for the completion of the task, DISTURBANCES. An exact analysis can be made in module Logistic Regression. However, the ultimate decision about which model to choose is up to the experiment maker.

18 DIMENSION REDUCTION AND GROUPING

As the number of variables subjected to a statistical analysis grows, their precision grows, but so does the level of complexity and difficulty in interpreting the obtained results. Too many variables increase the risk of their mutual correlation. The information carried by some variables can, then, be redundant, i.e. a part of the variables may not bring in new information for analysis but repeat the information already given by other variables. The need for dimension reduction (a reduction of the number of variables) has inspired a whole group of analyses devoted to that issue, such as: factor analysis, principal component analysis, or discriminant analysis. Those methods allow the detection of relationships among the variables. On the basis of those relationships one can distinguish, for further analysis, groups of similar variables and select only one representative (one variable) of each group, or a new variable the values of which are calculated on the basis of the remaining variables in the group. As a result, one can be certain that the information carried by each group is included in the analysis. In this manner we can reduce a set of variables p to a set of variables k where $k < p$, with only a small loss of information.

18.1 PRINCIPAL COMPONENT ANALYSIS

The window with settings for Principal component analysis is accessed via the menu Statistics → Multivariate Models → Principal Component Analysis.



Principal component analysis involves defining completely new variables (**principal components**) which are a linear combination of the observed (original) variables. An exact analysis of the principal components makes it possible to point to those original variables which have a big influence on the appearance of particular principal components, that is those variables which constitute a homogeneous group. A principal component is then a representative of that group. Subsequent components are mutually orthogonal (uncorrelated) and their number (k) is lower than or equal to the number of original variables (p).

Particular principal components are a linear combination of original variables:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{in}X_p$$

where:

X_1, X_2, \dots, X_p – original variables,
 $a_{i1}, a_{i2}, \dots, a_{ip}$ – coefficients of the i th principal component

Each principal component explains a certain part of the variability of the original variables. They are, then, naturally based on such measures of variability as covariance (if the original variables are of similar size and are expressed in similar units) or correlation (if the assumptions necessary in order to use covariance are not fulfilled).

Mathematical calculations which allow the distinction of principal components include defining the eigenvalues and the corresponding eigenvectors from the following matrix equation:

$$(M - \lambda I)a = 0$$

where:

λ – eigenvalues,
 $a_i = (a_{i1}, a_{i2}, \dots, a_{ip})$ – eigenvector corresponding to the i th eigenvalue,
 M – the variance matrix or covariance matrix of original variables X_1, X_2, \dots, X_p ,
 I – identity matrix (1 on the main diagonal, 0 outside of it).

18.1.1 The interpretation of coefficients related to the analysis

Every principal component is described by:

Eigenvalue

An eigenvalue informs about which part of the total variability is explained by a given principal component. The first principal component explains the greatest part of variance, the second principal component explains the greatest part of that variance which has not been explained by the previous component, and the subsequent component explains the greatest part of that variance which has not been explained by the previous components. As a result, each subsequent principal component explains a smaller and smaller part of the variance, which means that the subsequent values are smaller and smaller.

Total variance is a sum of the eigenvalues, which allows the calculation of the variability percentage defined by each component.

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_p} \cdot 100\%$$

Consequently, one can also calculate the cumulative variability and the cumulative variability percentage for the subsequent components.

Eigenvector

An eigenvector reflects the influence of particular original variables on a given principal component. It contains the $a_{i1}, a_{i2}, \dots, a_{ip}$ coefficients of a linear combination which defines a component. The sign of those coefficients points to the direction of the influence and is accidental which does not change the value of the carried information.

Factor loadings

Factor loadings, just as the coefficients included in the eigenvector, reflect the influence of particular variables on a given principal component. Those values illustrate the part of the variance of a given component is constituted by the original variables. When an analysis is based on the correlation matrix, we interpret those values as correlation coefficients between original variables and a given principal value.

Variable contributions

They are based on the determination coefficients between original variables and a given principal component. They show what percentage of the variability of a given principal component can be explained by the variability of particular original variables.

Communalities

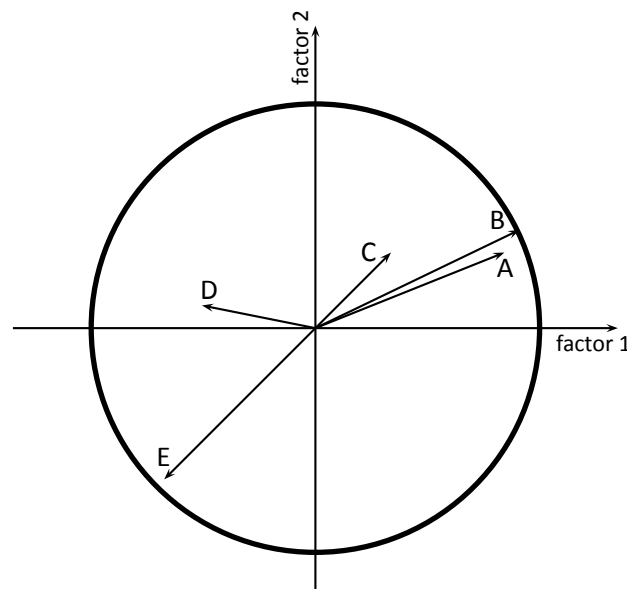
They are based on the determination coefficients between original variables and a given principal component. They show what percentage of a given original variable can be explained by the variability of a few initial principal components. For example: the result concerning the second variable contained in the column concerning the fourth principal component tells us what percent of the variability of the second variable can be explained by the variability of four initial principal components.

18.1.2 Graphical interpretation

A lot of information carried by the coefficients returned in the tables can be presented on one chart. The ability to read charts allows a quick interpretation of many aspects of the conducted analysis. The charts gather in one place the information concerning the mutual relationships among the components, the original variables, and the cases. They give a general picture of the principal components analysis which makes them a very good summary of it.

Factor loadings graph

The graph shows vectors connected with the beginning of the coordinate system, which represent original variables. The vectors are placed on a plane defined by the two selected principal components.



The coordinates of the terminal points of the vector are the corresponding factor loadings of the variables.

Vector length represents the information content of an original variable carried by the principal components which define the coordinate system. The longer the vector the greater the contribution of the original variable to the components. In the case of an analysis based on a correlation matrix the loadings are correlations between original variables and principal components. In such a case points fall into the unit circle. It happens because the correlation coefficient cannot exceed

one. As a result, the closer a given original variable lies to the rim of the circle the better the representation of such a variable by the presented principal components.

The sign of the coordinates of the terminal point of the vector i.e. the sign of the loading factor, points to the positive or negative correlation of an original variable and the principal components forming the coordination system. If we consider both axes (2 components) together then original variables can fall into one of four categories, depending on the combination of signs (+/–) and their loading factors.

The angle between vectors indicates the correlation of original values:

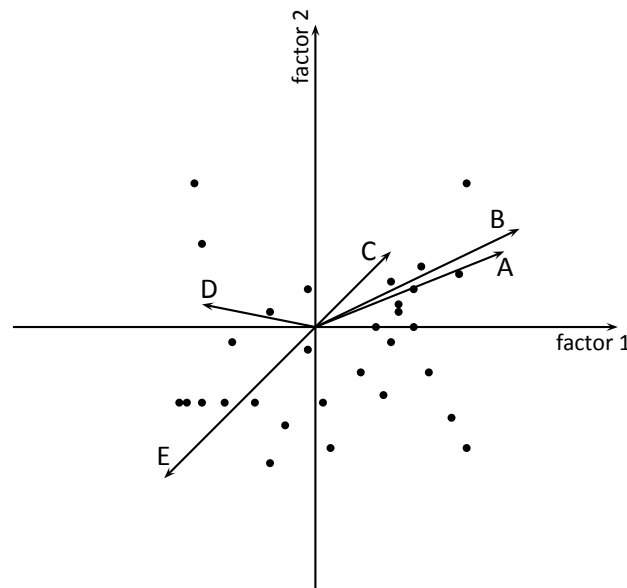
$0 < \alpha < 90^0$ – the smaller the angle between the vectors representing original variables, the stronger the positive correlation among these variables.

$\alpha = 90^0$ – the vectors are perpendicular, which means that the original variables are not correlated.

$90^0 < \alpha < 180^0$ – the greater the angle between the vectors representing the original variables, the stronger the negative correlation among these variables.

Biplot

The graph presents 2 series of data placed in a coordinate system defined by 2 principal components. The first series on the graph are data from the first graph (i.e. the vectors of original variables) and the second series are points presenting particular cases.



Point coordinates should be interpreted as standardized values, i.e. positive coordinates pointing to a value higher than the mean value of the principal component, negative ones to a lower value, and the higher the absolute value the further the points are from the mean. If there are untypical observations on the graph, i.e. outliers, they can disturb the analysis and should be removed, and the analysis should be made again.

The distances between the points show the similarity of cases: the closer (in the meaning of Euclidean distance) they are to one another, the more similar information is carried by the compared cases.

Orthographic projection of points on vectors are interpreted in the same manner as point coordinates, i.e. projections onto axes, but the interpretation concerns original variables and not principal

components. The values placed at the end of a vector are greater than the mean value of the original variable, and the values placed on the extension of the vector but in the opposite direction are values smaller than the mean.

18.1.3 The criteria of dimension reduction

There is not one universal criterion for the selection of the number of principal components. For that reason it is recommended to make the selection with the help of several methods.

The percentage of explained variance

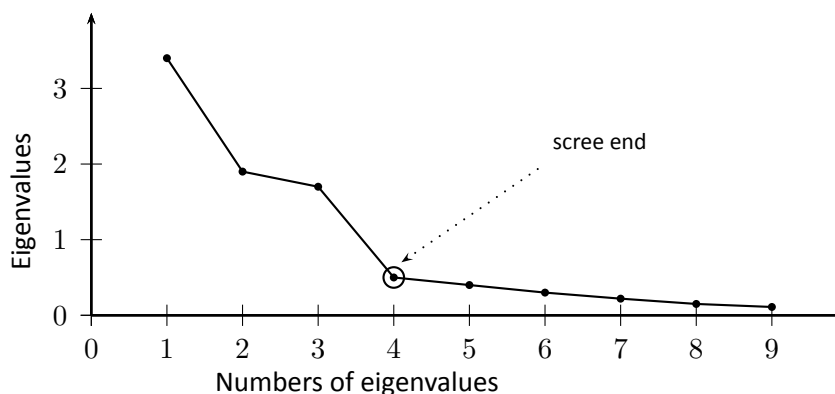
The number of principal components to be assumed by the researcher depends on the extent to which they represent original variables, i.e. on the variance of original variables they explain. All principal components explain 100% of the variance of original variables. If the sum of the variances for a few initial components constitutes a large part of the total variance of original variables, then principal components can satisfactorily replace original variables. It is assumed that the variance should be reflected in principal components to the extent of over 80 percent.

Kaiser criterion

According to the Kaiser criterion the principal components we want to leave for interpretation should have at least the same variance as any standardized original variable. As the variance of every standardized original variable equals 1, according to Kaiser criterion the important principal components are those the eigenvalue of which exceeds or is near value 1.

Scree plot

The graph presents the pace of the decrease of eigenvalues, i.e. the percentage of explained variance.



The moment on the chart in which the process stabilizes and the decreasing line changes into a horizontal one is the so-called end of the scree (the end of sprinkling of the information about the original values carried by principal components). The components on the right from the point which ends the scree represent a very small variance and are, for the most part, random noise.

18.1.4 Defining principal components

When we have decided how many principal components we need we can start generating them. In the case of principal components created on the basis of a correlation matrix they are computed as a linear combination of standardized original values. If, however, principal components have been created on the basis of a covariance matrix, they are computed as a linear combination of eigenvalues which have been centralized with respect to the mean of the original values.

The obtained principal components constitute new variables with certain advantages. First of all, the variables are not collinear. Usually there are fewer of them than original variables, sometimes much fewer, and they carry the same or a slightly smaller amount of information than the original values. Thus, the variables can easily be used in most multidimensional analyses.

18.1.5 The advisability of using the Principal component analysis

If the variables are not correlated (the Pearson's correlation coefficient is near 0), then there is no use to conduct a principal component analysis, as in such a situation every variable is already a separate component.

Bartlett's test

The test is used to verify the hypothesis that the correlation coefficients between variables are zero (i.e. the correlation matrix is an identity matrix).

Hypotheses:

$$\begin{aligned}\mathcal{H}_0 : M &= I, \\ \mathcal{H}_1 : M &\neq I.\end{aligned}$$

where:

M – the variance matrix or covariance matrix of original variables X_1, X_2, \dots, X_p ,

I – the identity matrix (1 on the main axis, 0 outside of it).

The test statistic has the form presented below:

$$\chi^2 = - \left(n - 1 - \frac{2p + 5}{6} \right) \sum_{i=1}^k \ln \lambda_i,$$

where:

p – the number of original variables,

n – size (the number of cases),

λ_i – i th eigenvalue.

That statistic has, asymptotically (for large expected frequencies), the distribution χ^2 with $p(p-1)/2$ degrees of freedom.

On the basis of [test statistics](#), [p value](#) is estimated and then compared with the significance level α :

$$\begin{aligned}\text{if } p \leq \alpha &\implies \text{ we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{ there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

The Kaiser-Meyer-Olkin coefficient

The coefficient is used to check the degree of correlation of original variables, i.e. the strength of the evidence testifying to the relevance of conducting a principal component analysis.

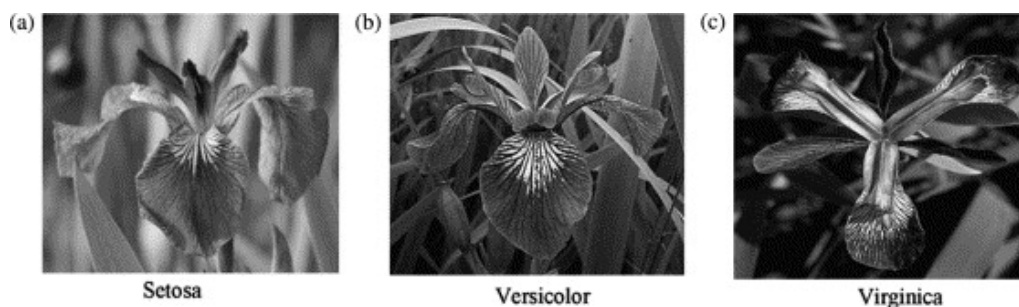
$$KMO = \frac{\sum_{i \neq j}^p \sum_{j \neq i}^p r_{ij}^2}{\sum_{i \neq j}^p \sum_{j \neq i}^p r_{ij}^2 + \sum_{i \neq j}^p \sum_{j \neq i}^p \hat{r}_{ij}^2},$$

r_{ij} – the correlation coefficient between the i th and the j th variable,

\hat{r}_{ij} – the partial correlation coefficient between the i th and the j th variable.

The value of the Kaiser coefficient belongs to the range $< 0, 1 >$ where low values testify to the lack of a need to conduct a principal component analysis, and high values are a reason for conducting such an analysis.

EXAMPLE 18.1. (file: iris.pqs) That classical set of data was first published in Ronald Aylmer Fisher's 1936[29] work in which discriminant analysis was presented. The file contains the measurements (in centimeters) of the length and width of the petals and sepals for 3 species of irises. The studied species are setosa, versicolor, and virginica. It is interesting how the species can be distinguished on the basis of the obtained measurements.



The photos are from scientific paper: Lee, et al. (2006r), "Application of a noisy data classification technique to determine the occurrence of flashover in compartment fires"

Principal component analysis will allow us to point to those measurements (the length and the width of the petals and sepals) which give the researcher the most information about the observed flowers.

The first stage of work, done even before defining and analyzing principal components, is checking the advisability of conducting the analysis. We start, then, from defining a correlation matrix of the variables and analyzing the obtained correlations with the use of Bartlett's test and the KMO coefficient.

Correlation matrix				
Variable	Sepal Len	Sepal Wic	Petal Len	Petal Wid
Sepal Len	1	-0,11757	0,871754	0,817941
Sepal Wic	-0,11757	1	-0,42844	-0,366126
Petal Len	0,871754	-0,42844	1	0,962865
Petal Wid	0,817941	-0,366126	0,962865	1

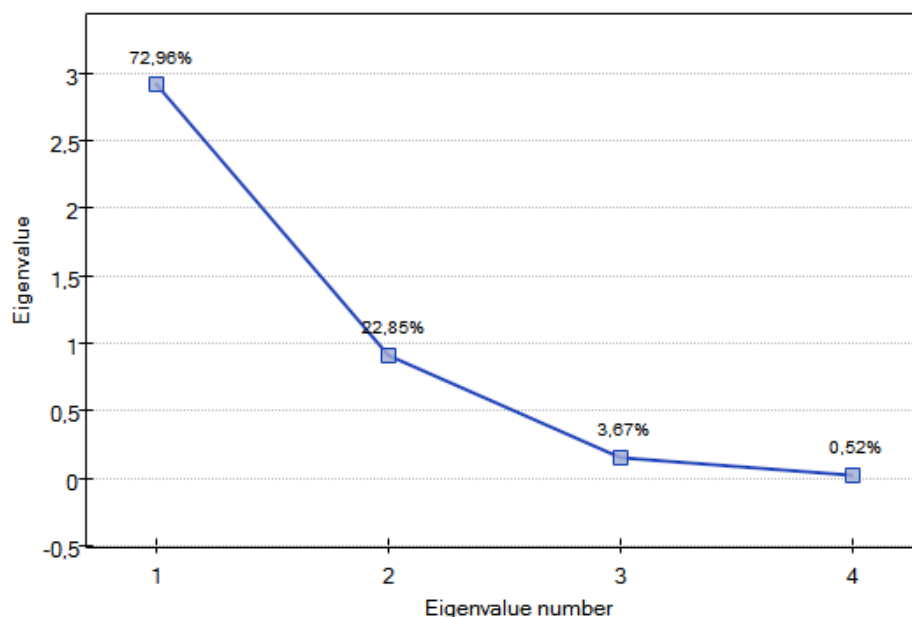
Principal Component Analysis	
Analysis time	0,57sec.
Analysed variables	Sepal Length;Sepal Width
Significance level	0,05
Analysis of correlation matrix	
Bartlett test	
Chi-square statistic	706,959243
Degrees of freedom	6
p-value	<0.000001
Kaiser-Mayer-Olkin coefficient	
KMO	0,540077

The value p of Bartlett's statistics points to the truth of the hypothesis that there is a significant difference between the obtained correlation matrix and the identity matrix, i.e. that the data are strongly correlated. The obtained KMO coefficient is average and equals 0.54. We consider the indications for conducting a principal component analysis to be sufficient.

The first result of that analysis which merits our special attention are eigenvalues:

Eigenvalues				
Number	Eigenvalu	% varianc	Cumulativ	% cumula
1	2,918498	72,96244	2,918498	72,96244
2	0,91403	22,85076	3,832528	95,81320
3	0,146757	3,668922	3,979285	99,48212
4	0,020715	0,517871	4	100

The obtained eigenvalues show that one or even two principal components will describe our data well. The eigenvalue of the first component is 2.92 and the percent of the explained variance is 72.96. The second component explains much less variance, i.e. 22.85%, and its eigenvalue is 9.91. According to Kaiser criterion, one principal component is enough for an interpretation, as only for the first principal component the eigenvalue is greater than 1. However, looking at the graph of the scree we can conclude that the decreasing line changes into a horizontal one only at the third principal component.



From that we may infer that the first two principal components carry important information. Together they explain a great part, as much as 95.81%, of the variance (see the cumulative % column).

The communalities for the first principal component are high for all original variables except the variable of the width of the sepal, for which they equal 21.17%. That means that if we only interpret the first principal component, only a small part of the variable of the width of the sepal would be reflected.

% communalities				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Len	79,24004	92,25986	99,85857	100
Sepal Wid	21,17313	99,09193	99,9684	100
Petal Len	98,31816	98,37299	98,66944	100
Petal Wid	93,11843	93,52803	99,43209	100

For the first two principal components the communalities are at a similar, very high level and they exceed 90% for each of the analyzed variables, which means that with the use of those components the variance of each variability is represented in over 90%.

In the light of all that knowledge it has been decided to separate and interpret 2 components.

In order to take a closer look at the relationship of principal components and original variables, that

is the length and the width of the petals and sepals, we interpret: eigenvectors, factor loadings, and contributions of original variables.

Eigenvectors				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Len	-0,52106	-0,37741	0,719566	0,261286
Sepal Wid	0,269347	-0,92329	-0,24438	-0,12351
Petal Len	-0,58041	-0,02449	-0,14212	-0,80144
Petal Wid	-0,56485	-0,06694	-0,63427	0,523597

Factor loadings				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Len	-0,89016	-0,36083	0,275658	0,037606
Sepal Wid	0,460143	-0,88271	-0,09362	-0,01777
Petal Len	-0,99155	-0,02341	-0,05444	-0,11535
Petal Wid	-0,96497	-0,064	-0,24298	0,07536

% variable contributions				
Variable	Factor1	Factor2	Factor3	Factor4
Sepal Len	27,15096	14,24440	51,77757	6,827052
Sepal Wid	7,254804	85,24748	5,972245	1,525463
Petal Len	33,68793	0,059984	2,01999	64,23208
Petal Wid	31,90629	0,448123	40,23019	27,41539

Particular original variables have differing effects on the first principal component. Let us put them in order according to that influence:

1. The length of a petal is negatively correlated with the first component, i.e. the longer the petal, the lower the values of that component. The eigenvector of the length of the petal is the greatest in that component and equals -0.58. Its factor loading informs that the correlation between the first principal component and the length of the petal is very high and equals -0.99 which constitutes 33.69% of the first component;
2. The width of the petal has an only slightly smaller influence on the first component and is also negatively correlated with it;
3. We interpret the length of the sepal similarly to the two previous variables but its influence on the first component is smaller;
4. The correlation of the width of the sepal and the first component is the weakest, and the sign of that correlation is positive.

The second component represents chiefly the original variable "sepal width"; the remaining original variables are reflected in it to a slight degree. The eigenvector, factor loading, and the contribution of the variable "sepal width" is the highest in the second component.

Each principal component defines a homogeneous group of original values. We will call the first component "petal size" as its most important variables are those which carry the information about the petal, although it has to be noted that the length of the sepal also has a significant influence on the value of that component. When interpreting we remember that the greater the values of that component, the smaller the petals.

We will call the second component "sepal width" as only the width of the sepal is reflected to a greater degree here. The greater the values of that component, the narrower the sepal.

Finally, we will generate the components by choosing, in the analysis window, the option: Add Principal Components. A part of the obtained result is presented below:

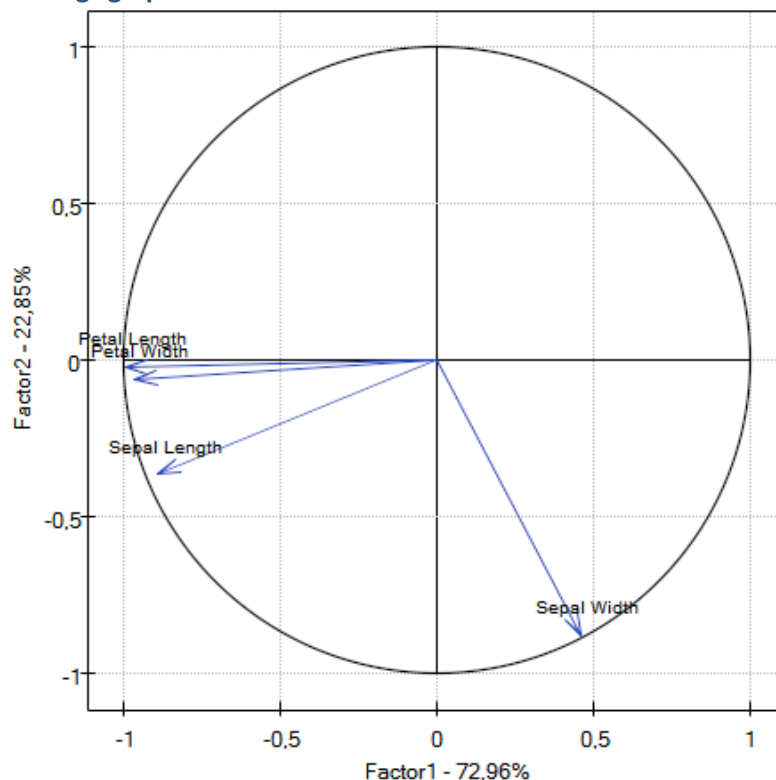
Principal Components			
Factor1	Factor2	Factor3	Factor4
2,257141	-0,47842	0,12728	0,024088
2,074013	0,671883	0,233826	0,102663
2,356335	0,340766	-0,04405	0,028282
2,291707	0,5954	-0,09098	-0,06573
2,381863	-0,64467	-0,01568	-0,03580
2,068701	-1,48420	-0,02687	0,006586
2,435868	-0,04748	-0,33435	-0,03665
2,225392	-0,22240	0,088399	-0,02453
2,326845	1,111604	-0,14459	-0,02677
2,177035	0,467448	0,252918	-0,03976
2,159077	-1,04020	0,267784	0,016676
2,318364	-0,13263	-0,09344	-0,13303
2,211044	0,726243	0,23014	0,002417
2,624309	0,958296	-0,18019	-0,01915
2,191399	-1,85384	0,471322	0,194082

In order to be able to use the two initial components instead of the previous four original values, we copy and paste them into the data sheet. Now, the researcher can conduct the further statistics on two new, uncorrelated variables.

Analysis of the graphs of the two initial components

The analysis of the graphs not only leads the researcher to the same conclusions as the analysis of the tables but will also give him or her the opportunity to evaluate the results more closely.

Factor loadings graph



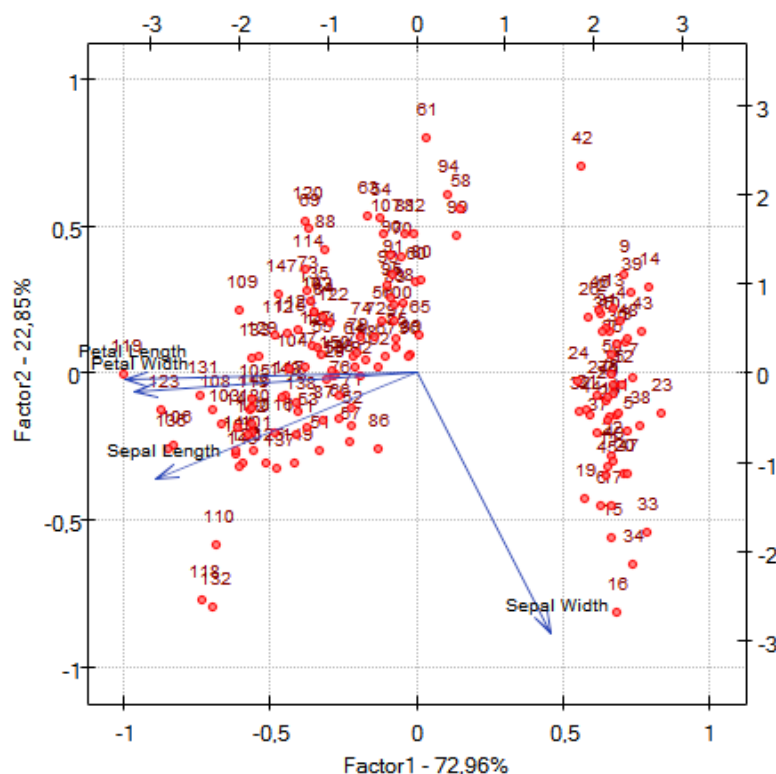
The graph shows the two first principal components which represent 72.96% of the variance and 22.85% of the variance, together amounting to 95.81% of the variance of original values

The vectors representing original values almost reach the rim of the unit circle (a circle with the radius of 1), which means they are all well represented by the two initial principal components which form the coordinate system.

The angle between the vectors illustrating the length of the petal, the width of the petal, and the length of the sepal is small, which means those variables are strongly correlated. The correlation of those variables with the components which form the system is negative, the vectors are in the third quadrant of the coordinate system. The observed values of the coordinates of the vector are higher for the first component than for the second one. Such a placement of vectors indicates that they comprise a uniform group which is represented mainly by the first component.

The vector of the width of the sepal points to an entirely different direction. It is only slightly correlated with the remaining original values, which is shown by the inclination angle with respect to the remaining original values – it is nearly a right angle. The correlation of that vector with the first component is positive and not very high (the low value of the first coordinate of the terminal point of the vector), and it is negative and high (the high value of the second coordinate of the terminal point of the vector) in the case of the second component. From that we may infer that the width of the sepal is the only original variable which is well represented by the second component.

Biplot



The biplot presents two series of data spread over the first two components. One series are the vectors of original values which have been presented on the previous graph and the other series are the points which carry the information about particular flowers. The values of the second series are read on the upper axis X and the right axis Y . The manner of interpretation of vectors, that is the first series, has been discussed with the previous graph. In order to understand the interpretation of points let us focus on flowers number 33, 34, and 109.

Flowers number 33 and 34 are similar – the distance between points 33 and 34 is small. For both points the value of the first component is much greater than the average and the value of

the second component is much smaller than the average. The average value, i.e. the arithmetic mean of both components, is 0, i.e. it is the middle of the coordination system. Remembering that the first component is mainly the size of the petals and the second one is mainly the width of the sepal we can say that flowers number 33 and 34 have small petals and a large width of the sepal. Flower number 109 is represented by a point which is at a large distance from the other two points. It is a flower with a negative first component and a positive, although not high second component. That means the flower has relatively large petals while the width of the sepal is a bit smaller than average.

Similar information can be gathered by projecting the points onto the lines which extend the vectors of original values. For example, flower 33 has a large width of the sepal (high and positive values on the projection onto the original value "sepal width") but small values of the remaining original values (negative values on the projection onto the extension of the vectors illustrating the remaining original values).

19 SURVIVAL ANALYSIS

Survival analysis is often used in medicine. In other fields of study it is also called reliability analysis, duration analysis, or event history analysis. Its main goal is to evaluate the remaining time of the survival of, for example, patients after an operation. The tools used in the analysis are life tables and Kaplan-Meier curves. Another interesting aspect of that issue is comparing the survival time of, for example, patients treated according to different protocols. For that purpose comparisons of two or more survival curves are used. A number of methods (regression models) have also been created for studying the influence of various variables on the survival time. In order to make the understanding of the issue easier, the example of the length of the life of patients after a heart transplantation will be used to illustrate basic definitions.

Event – is the change interesting to the researcher, e.g. death;

Survival time – is the period of time between the initial state and the occurrence of a given event, e.g. the length of a patient's life after a heart transplantation.

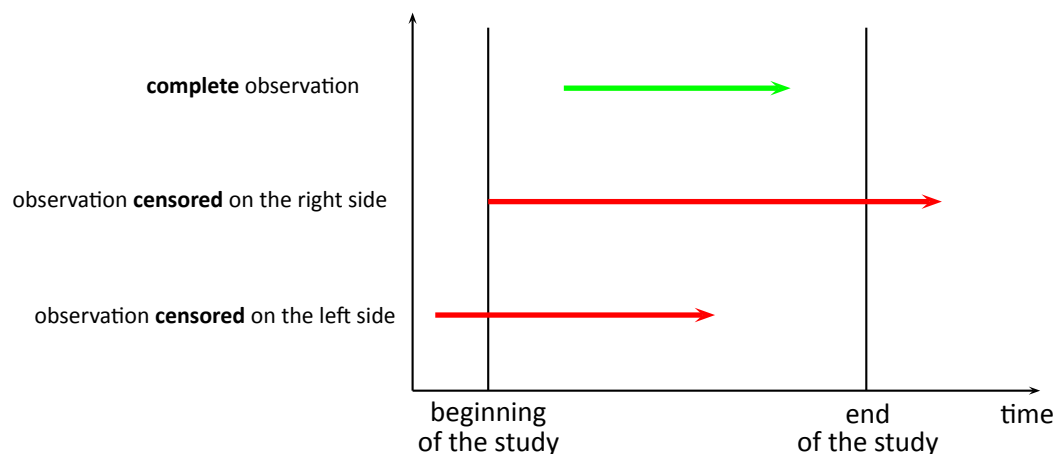
Note!

In the analysis one column with the calculated time ought to be marked. When we have at our disposal two points in time: the initial and the final ones, before the analysis we calculate the time between the two points, using the datasheet formulas.

Censored observations – are the observations for which we only have incomplete pieces of information about the survival time.

Censored and complete observations – an example concerning the survival time after a heart transplantation:

- **a complete observation** – we know the date of the transplantation and the date of the patient's death so we can establish the exact survival time after the transplantation.
- **observation censored on the right side** – the date of the patient's death is not known (the patient is alive when the study finishes) so the exact survival time cannot be established.
- **observation censored on the left side** – the date of the heart transplantation is not known but we know it was before this study started, and we cannot establish the exact survival time.

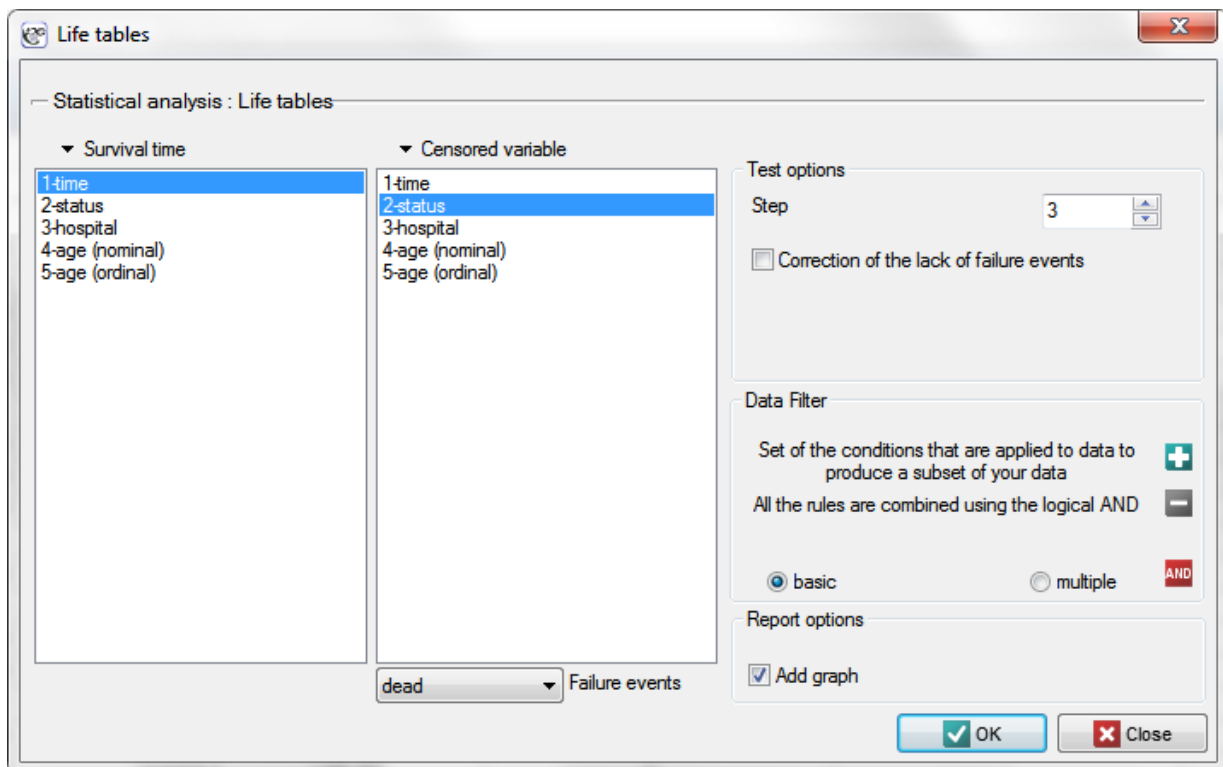


Note!

The end of the study means the end of the observation of the patient. It is not always the same moment for all patients. It can be the moment of losing touch with the patient (so we do not now the patient's survival time). Analogously, the beginning of the study does not have to be the same point in time for all patients.

19.1 LIFE TABLES

The window with settings for life tables is accessed via the menu Statistics→Survival analysis→Life tables



Life tables are created for time ranges with equal spans, provided by the researcher. The ranges can be defined by giving the step. For each range PQStat calculates:

- **the number of entered cases** — the number of people who survived until the time defined by the range;
- **the number of censored cases** — the number of people in a given range qualified as censored cases;
- **the number of cases at risk** — the number of people in a given range minus a half of the censored cases in the given range;
- **the number of complete cases** — the number of people who experienced the event (i.e. died) in a given range;
- **proportions of complete cases** — the proportion of the number of complete cases (deaths) in a given range to the number of the cases at risk in that range;
- **proportions of the survival cases** — calculated as 1 minus the proportion of complete cases in a given range;

- **cumulative survival proportion (survival function)** — the probability of surviving over a given period of time. Because to survive another period of time, one must have survived all the previous ones, the probability is calculated as the product of all the previous proportions of the survival cases.

± standard error of the survival function;

- **probability density** — the calculated probability of experiencing the event (death) in a given range, calculated in a period of time;

± standard error of the probability density;

- **hazard rate** — probability (calculated per a unit of time) that a patient who has survived until the beginning of a given range will experience the event (die) in that range;

± standard error of the hazard rate

Note!

In the case of a lack of complete observations in any range of survival time range there is the possibility of using correction. The zero number of complete cases is then replaced with value 0.5.

Graphic interpretation

We can illustrate the information obtained thanks to the life tables with the use of several charts:

- a survival function graph,
- a probability density graph,
- a hazard rate graph.

EXAMPLE 19.1. (file: transplant.pqs)

Patients' survival rate after the transplantation of a liver was studied. 89 patients were observed over 21 years. The age of a patient at the time of the transplantation was in the range of (45years; 60years). A fragment of the collected data is presented in the table below:

time	status	hospital	age (nominal)	age (ordinal)
14	dead	hospital 1	<45; 50)	1
21	alive	hospital 1	<45; 50)	1
4	dead	hospital 1	<50; 55)	2
4	alive	hospital 1	<50; 55)	2
5	dead	hospital 1	<50; 55)	2
6	alive	hospital 1	<50; 55)	2
6	alive	hospital 1	<50; 55)	2
9	dead	hospital 1	<50; 55)	2
16	alive	hospital 1	<50; 55)	2

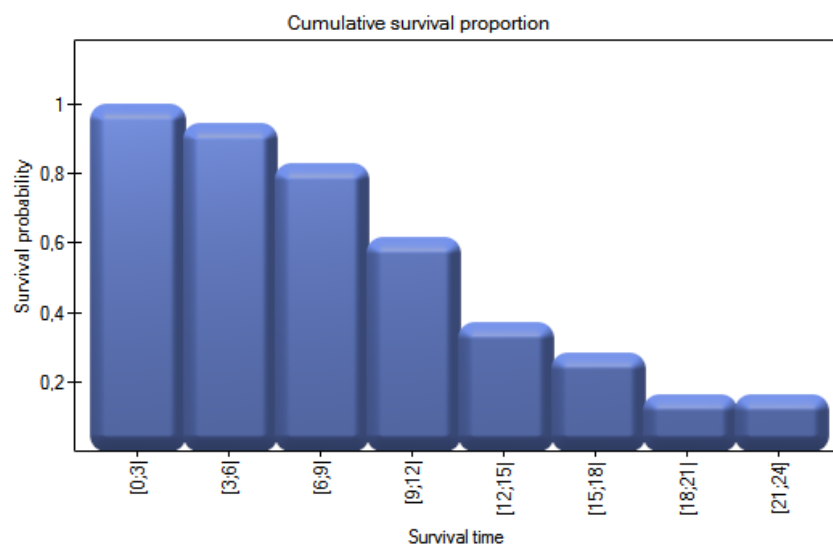
The complete data in the analysis are those as to which we have complete information about the length of life after the transplantation, i.e. described as "death" (it concerns 53 people which constitutes 59.55% of sample). The censored data are those about which we do not have that information because at the time when the study was finished the patients were alive (36 people, i.e. 40.45% of them). We build the life tables of those patients by creating time periods of 3 years:

Life tables												
Interval	Censored	Failure eve	Entered	At risk	Failure eve	Censored	Cumulative Probability	Hazard rat	Std. error	Std. error	Std. error	Std. error
[0;3]	0	5	89	89	0,0561797	0,9438202	1	0,0187265	0,0192678	0	0,0081361	0,0086132
[3;6]	5	10	84	81,5	0,1226993	0,8773006	0,9438202	0,0386020	0,0435729	0,0244084	0,0114771	0,0137495
[6;9]	14	16	69	62	0,2580645	0,7419354	0,8280140	0,0712270	0,0987654	0,0404362	0,0157274	0,0244188
[9;12]	7	14	39	35,5	0,3943661	0,6056338	0,6143330	0,0807573	0,1637426	0,0549303	0,0182830	0,0424215
[12;15]	3	4	18	16,5	0,2424242	0,7575757	0,3720608	0,0300655	0,0919540	0,0603812	0,0139645	0,0455375
[15;18]	3	4	11	9,5	0,4210526	0,5789473	0,2818642	0,0395598	0,1777777	0,0602764	0,0172650	0,0856701
[18;21]	2	0	4	3	0	1	0,1631845	0	0	0,0570648	0	0
[21;24]	2	0	2	1	0	1	0,1631845			0,0570648		

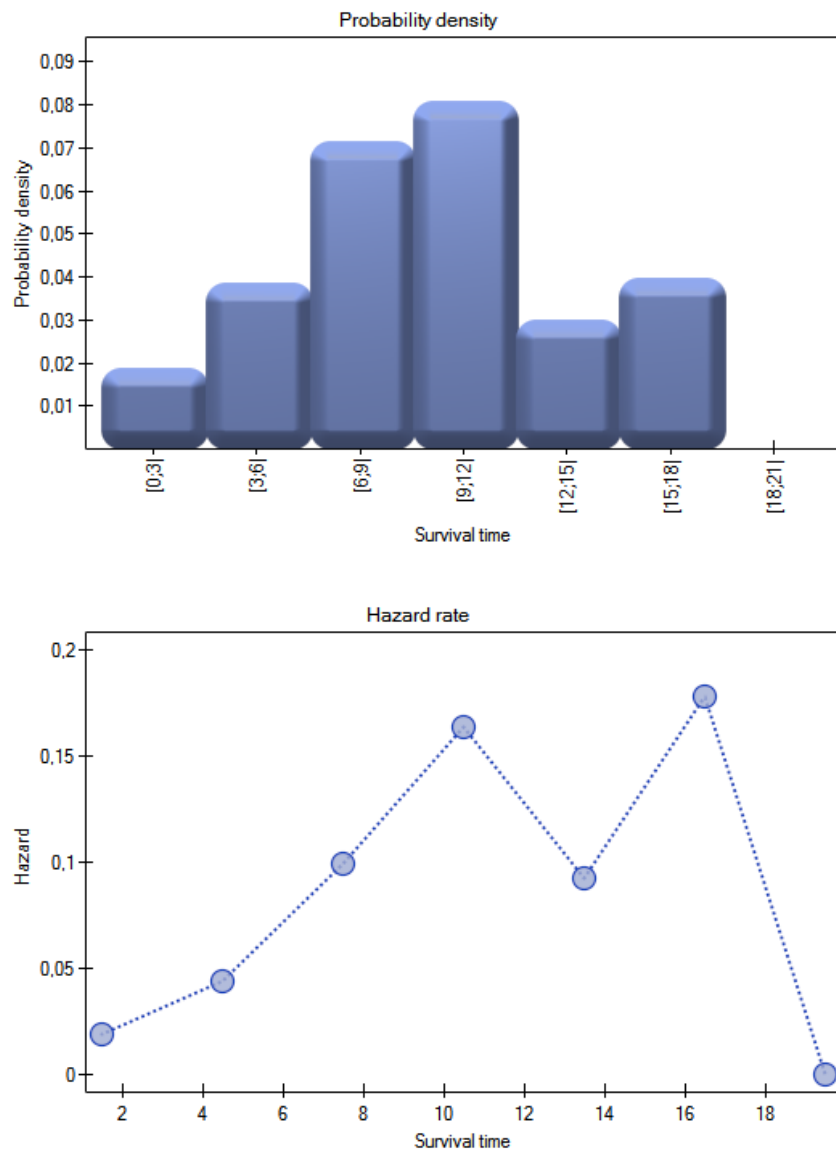
For each 3-year period of time we can interpret the results obtained in the table, for example, for people living for at least 9 years after the transplantation who are included in the range [9;12]:

- the number of people who survived 9 years after the transplantation is 39,
- there are 7 people about whom we know they had lived at least 9-12 years at the moment the information about them was gathered but we do not know if they lived longer as they were left out of the study after that time,
- the number of people at the risk of death in that age range is 36,
- there are 14 people about whom we know they died 9 to 12 years after the transplantation,
- 39.4% of the endangered patients died 9 to 12 years after the transplantation,
- 60.6% of the endangered patients lived 9 to 12 years after the transplantation,
- the percent of survivors 9 years after the transplantation is $61.4\% \pm 5\%$,
- $0,08 \pm 0.02$ is the death probability for each year from the 9-12 range.

The results will be presented on a few graphs:



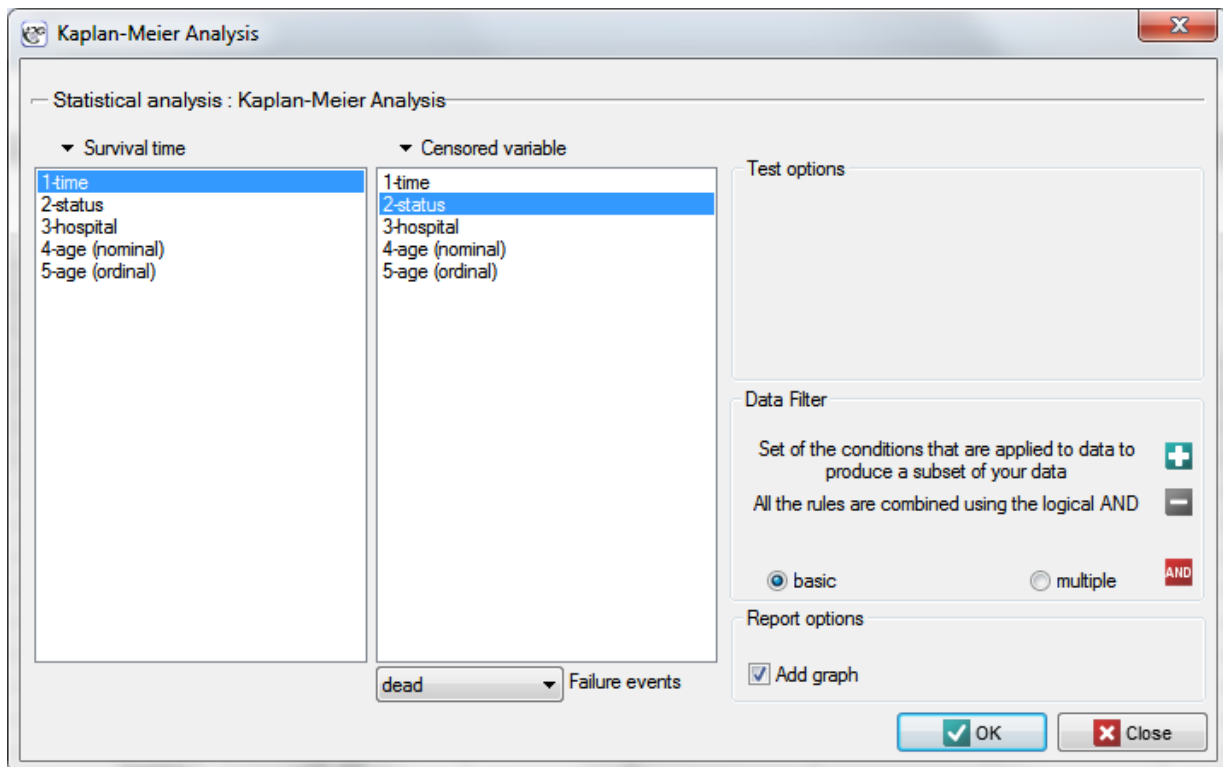
The probability of survival decreases with the time passed since the transplantation. We do not, however, observe a sudden plunge of the survival function, i.e. a period of time in which the probability of death would rise dramatically.



19.2 KAPLAN-MEIER CURVES

Kaplan-Meier curves allow the evaluation of the survival time without the need to arbitrarily group the observations like in the case of life tables. The estimator was introduced by Kaplan and Meier (1958)[41].

The window with settings for Kaplan-Meier curve is accessed via the menu Survival analysis→ Multi-dimensional Models→Kaplan-Meier Analysis

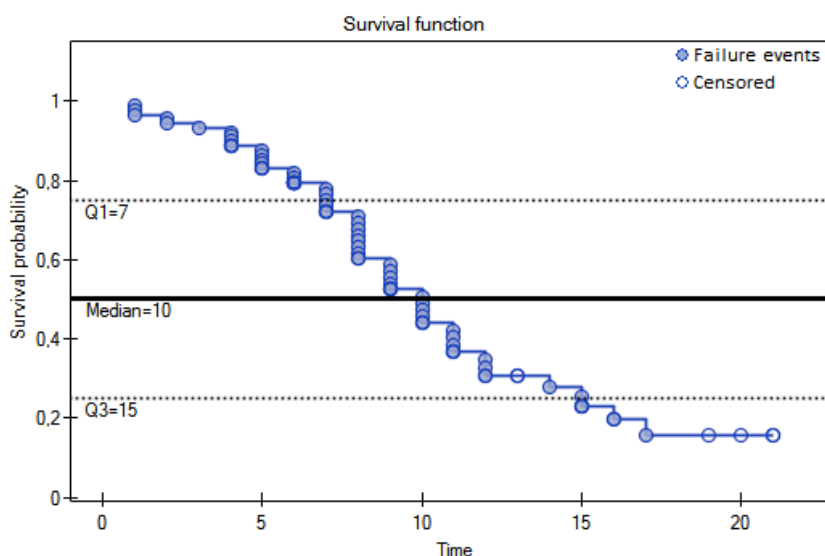


As with survival tables we calculate the survival function, i.e. the probability of survival until a certain time. The graph of the Kaplan-Meier survival function is created by a step function. The point of time at which the value of the function is 0.5 is the **survival time median**. That is the time of the observation below which half of the observed patients have died and half of them are still alive. Both the median and other percentiles are determined as the shortest survival time for which the survival function is smaller or equal to a given percentile. **The survival time mean** is determined as the field under the survival curve.

The data concerning the survival time are usually very heavily skewed so in the survival analysis the median is a better measure of the central tendency than the mean.

Example (19.1) continued (file: *transplant.pqs*)

We present the survival time after a liver transplantation, with the use of the Kaplan-Meier curve.



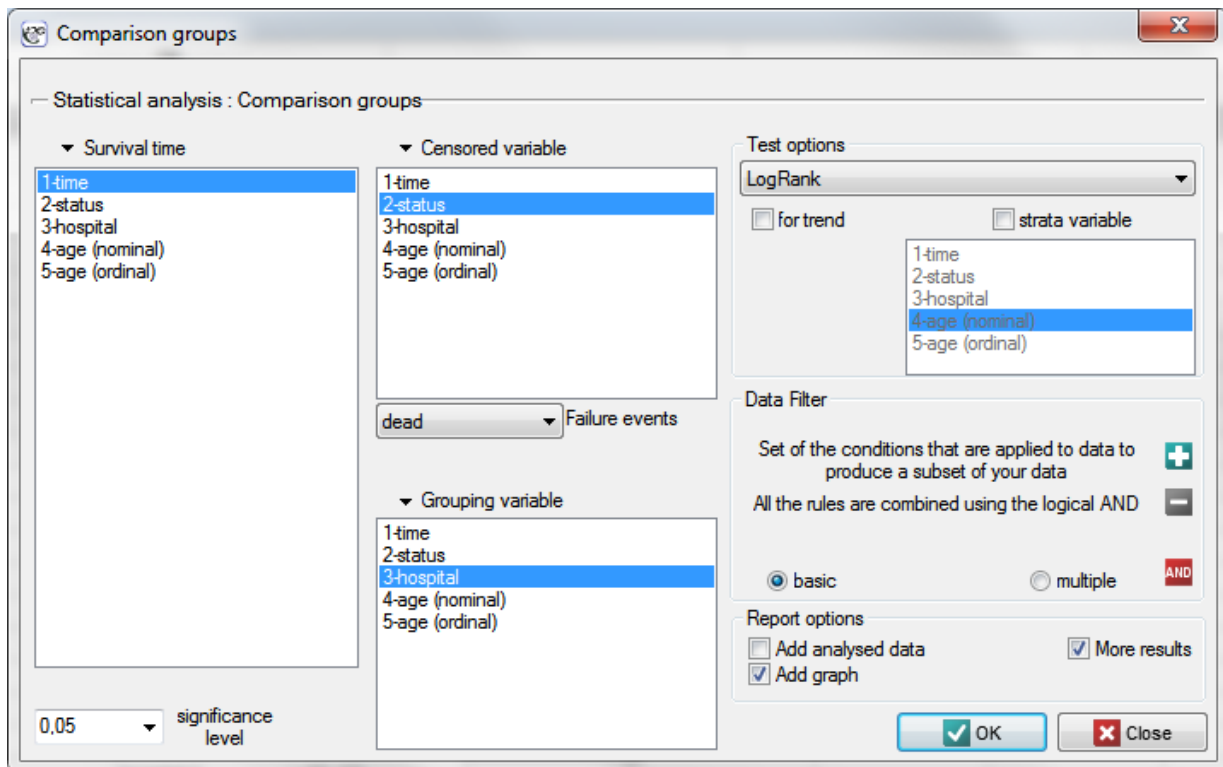
Kaplan-Meier Analysis	
Analysis time	0,12sec.
Analysed variables	time;status
Censored variable	status(dead;alive)
Frequency	89
Failure events	dead
Frequency	53
Percent	59,55%
Censored	alive
Frequency	36
Percent	40,45%
Survival time	
Lower quartile	7
Median	10
Upper quartile	15
Mean	10,954902442

The survival function does not suddenly plunge right after the transplantation. Therefore, we conclude that the initial period after the transplantation does not carry a particular risk of death. The value of the median shows that for 10 years after the transplantation a half of the patients have died and another half is still alive. The value is marked on the graph by drawing a line in point 0.5 which signifies the median. In a similar manner we mark the quartiles in the graph.

19.3 COMPARISON OF SURVIVAL CURVES

The survival functions can be built separately for different subgroups, e.g. separately for women and men, and then compared. Such a comparison may concern two curves or more.

The window with settings for the comparison of survival curves is accessed via the menu Statystyka→Survival analysis→Comparison groups



Comparisons of k survival curves S_1, S_2, \dots, S_k , at particular points of the survival time t , in the program can be made with the use of three tests:

Log-rank test the most popular test drawing on the Mantel-Heanszel procedure for many 2×2 tables (Mantel-Heanszel 1959[56], Mantel 1966[58], Cox 1972[23]),

Gehan's generalization of Wilcoxon's test deriving from Wilcoxon's test (Breslow 1970, Gehan 1965[34][35]),

Tarone-Ware test deriving from Wilcoxon's test (Tarone and Ware 1977[76]).

The three tests are based on the same test statistic, they only differ in **weights** w_j the particular points of the timeline on which the test statistic is based.

Log-rank test: $w_j = 1$ – all the points of the timeline have the same weight which gives the later values of the timeline a greater influence on the result;

Gehan's generalization of Wilcoxon's test: $w_j = n_j$ – time moments are weighted with the number of observations in each of them, so greater weights are ascribed to the initial values of the time line;

Tarone-Ware test: $w_j = \sqrt{n_j}$ – time moments are weighted with the root of the number of observations in each of them, so the test is situated between the two tests described earlier.

An important condition for using the tests above is the proportionality of hazard. Hazard, defined as the slope of the survival curve, is the measure of how quickly a failure event takes place. Breaking the principle of hazard proportionality does not completely disqualify the tests above but it carries some risks. First of all, the placement of the point of the intersection of the curves with respect to the timeline has a decisive influence on decreasing the power of particular tests.

19.3.1 Differences among the survival curves

Hypotheses:

$$\mathcal{H}_0 : S_1(t) = S_2(t) = \dots = S_k(t), \quad \text{for all } t,$$

$$\mathcal{H}_1 : \text{not all } S_i(t) \text{ are equal.}$$

In calculations was used chi-square statistics form:

$$\chi^2 = U'V^{-1}U$$

where:

$$U_i = \sum_{j=1}^m w_j (d_{ij} - e_{ij})$$

V - covariance matrix of dimensions $(k-1) \times (k-1)$

where:

$$\text{diagonal: } \sum_{j=1}^m w_j^2 \frac{n_{ij}(n_j - n_{ij})d_j(n_j - d_j)}{n_j^2(n_j - 1)},$$

$$\text{off diagonal: } \sum_{j=1}^m w_j^2 \frac{n_{ij}n_{lj}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

m – number of moments in time with failure event (death),

$d_j = \sum_{i=1}^k d_{ij}$ – observed number of failure events (deaths) in the j -th moment of time,

d_{ij} – observed number of failure events (deaths) in the w i -th group w in the j -th moment of time,

$e_{ij} = \frac{n_{ij}d_j}{n_j}$ – expected number of failure events (deaths) in the w i -th group w in the j -th moment of time,

$n_j = \sum_{i=1}^k n_{ij}$ – the number of cases at risk in the j -th moment of time.

The statistic asymptotically (for large sizes) has the χ^2 distribution with $df = k - 1$ degrees of freedom.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\text{if } p \leq \alpha \implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1,$$

$$\text{if } p > \alpha \implies \text{there is no reason to reject } \mathcal{H}_0.$$

Hazard ratio

In the log-rank test the observed values of failure events (deaths) $O_i = \sum_{j=1}^m d_{ij}$ and the appropriate expected values $E_i = \sum_{j=1}^m e_{ij}$ are given.

The measure for describing the size of the difference between a pair of survival curves is the hazard ratio (HR).

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

If the hazard ratio is greater than 1, e.g. $HR = 2$, then the degree of the risk of a failure event in the first group is twice as big as in the second group. The reverse situation takes place when HR is smaller than one. When HR is equal to 1 both groups are equally at risk.

Note!

The confidence interval for HR is calculated on the basis of the standard deviation of the HR logarithm (Armitage and Berry 1994[5]).

19.3.2 Survival curve trend

Hypotheses:

\mathcal{H}_0 : In the studied population there is no trend in the placement of the S_1, S_2, \dots, S_k curves,

\mathcal{H}_1 : In the studied population there is a trend in the placement of the S_1, S_2, \dots, S_k curves.

In the calculation the chi-square statistic was used, in the following form:

$$\chi^2 = \frac{(c'U)^2}{c'Vc}$$

where:

$c = (c_1, c_2, \dots, c_k)$ – vector of the weights for the compared groups, informing about their natural order (usually the subsequent natural numbers).

The statistic asymptotically (for large sizes) has the χ^2 distribution with 1 degree of freedom.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

In order to conduct a trend analysis in the survival curves the grouping variable must be a numerical variable in which the values of the numbers inform about the natural order of the groups. The numbers in the analysis are treated as the c_1, c_2, \dots, c_k weights.

19.3.3 Survival curves for the stratas

Often, when we want to compare the survival times of two or more groups, we should remember about other factors which may have an impact on the result of the comparison. An adjustment (correction) of the analysis by such factors can be useful. For example, when studying rest homes and comparing the length of the stay of people below and above 80 years of age, there was a significant difference in the results. We know, however, that sex has a strong influence on the length of stay and the age of the inhabitants of rest homes. That is why, when attempting to evaluate the impact of age, it would be a good idea to stratify the analysis with respect to sex.

Hypotheses for the differences in survival curves:

$$\begin{aligned} \mathcal{H}_0 : S_1^*(t) &= S_2^*(t) = \dots = S_k^*(t), \quad \text{for all } t, \\ \mathcal{H}_1 : \text{not all } S_i^*(t) &\text{ are equal.} \end{aligned}$$

Hypotheses for the analysis of trends in survival curves:

\mathcal{H}_0 : In the studied population there is no trend in the placement of the $S_1^*, S_2^*, \dots, S_k^*$ curves,

\mathcal{H}_1 : In the studied population there is a trend in the placement of the $S_1^*, S_2^*, \dots, S_k^*$ curves.

where $S_1^*(t), S_2^*(t), \dots, S_k^*(t)$ -are the survival curves after the correction by the variable determining the strata.

The calculations for test statistics are based on formulas described for the tests, not taking into account the strata, with the difference that matrix U and V is replaced with the sum of matrices $\sum_{l=1}^L U$ and $\sum_{l=1}^L V$. The summation is made according to the strata created by the variables with respect to which we adjust the analysis $l=1, 2, \dots, L$

The statistic asymptotically (for large sizes) has the χ^2 distribution with 1 degree of freedom.

On the basis of test statistics, p value is estimated and then compared with the significance level α :

if $p \leq \alpha \implies$ we reject \mathcal{H}_0 and accept \mathcal{H}_1 ,
 if $p > \alpha \implies$ there is no reason to reject \mathcal{H}_0 .

Example (19.1) continued (*file transplant.pqs*)

The differences for two survival curves

Liver transplantations were made in two hospitals. We will check if the patients' survival time after transplantations depended on the hospital in which the transplantations were made. The comparisons of the survival curves for those hospitals will be made on the basis of all tests proposed in the program for such a comparison.

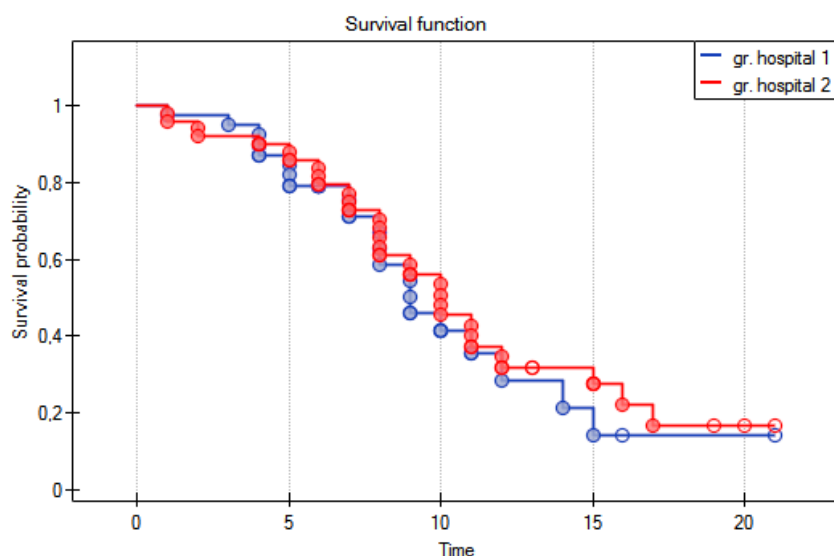
Hypotheses:

\mathcal{H}_0 : the survival curve of the patients of hospital no. 1 = the survival curve of the patients of hospital no. 2,
 \mathcal{H}_1 : the survival curve of the patients of hospital no. 1 \neq the survival curve of the patients of hospital no. 2.

Comparison groups	
Analysis time	0,17sec.
Analysed variables	time;status
Significance level	0,05
Grouping variable	hospital(hospital 1;hospital 2
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Chi-square statistic	0,274351308
Degrees of freedom	1
p-value	0,600427701

Logrank			
Group	Obs.	Exp.	Obs./Exp.
hospital 1	21	19,257088	1,0905075
hospital 2	32	33,742911	0,9483473

Logrank			
Group	Hazard r.	-95%CI	+95%CI
hospital 1<	1,1499031	0,6569948	2,0126144



On the basis of the significance level $\alpha = 0.05$, based on the obtained value $p=0.6004$ for the log-rank test ($p=0.6959$ for Gehan's and 0.6465 for Tarone-Ware) we conclude that there is no basis for rejecting the hypothesis \mathcal{H}_0 . The length of life calculated for the patients of both hospitals is similar.

The same conclusion will be reached when comparing the risk of death for those hospitals by determining the risk ratio. The obtained estimated value is $HR = 1.1499$ and 95% of the confidence interval for that value contains 1: $\langle 0.6570, 2.0126 \rangle$.

Differences for many survival curves

Liver transplantations were made for people at different ages. 3 age groups were distinguished: $\langle 45$ years; 50 years), $\langle 50$ years; 55 years), $\langle 55$ years; 60 years). We will check if the patients' survival time after transplantations depended on their age at the time of the transplantation.

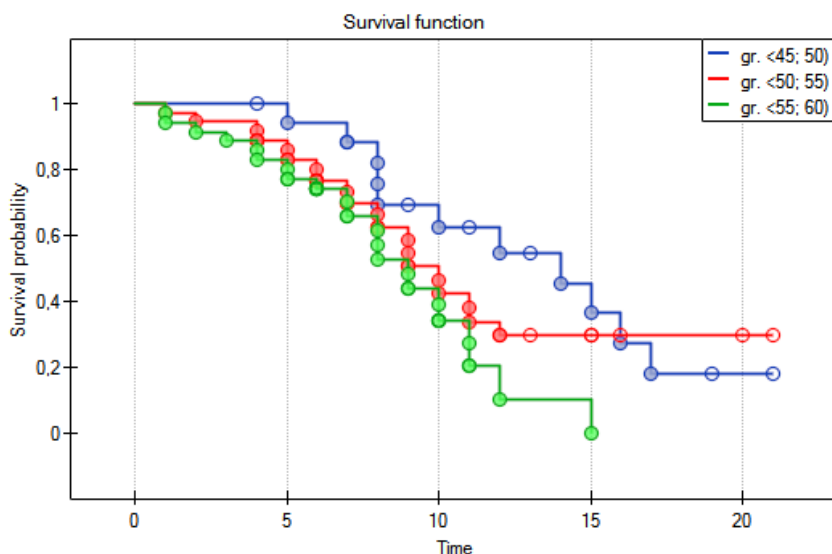
Hypotheses:

- \mathcal{H}_0 : survival rates of patients aged $\langle 45$ years; 50 years), $\langle 50$ years; 55 years), $\langle 55$ years; 60 years) are similar,
 \mathcal{H}_1 : at least one survival curve out of the 3 curves above differs from the other curves.

Comparison groups	
Analysis time	0,18sec.
Analysed variables	time;status
Significance level	0,05
Grouping variable	age (nominal)(<45; 50); <50
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Chi-square statistic	5,342329381
Degrees of freedom	2
p-value	0,069171615

Logrank			
Group	Obs.	Exp.	Obs./Exp.
<45; 50)	11	16,120092	0,6823782
<50; 55)	20	21,490397	0,9306482
<55; 60)	22	15,389510	1,4295451

Logrank			
Group	Hazard r.	-95%CI	+95%CI
<45; 50)<	0,7332289	0,3843909	1,3986403
<45; 50)<	0,4773394	0,2373928	0,9598135
<50; 55)<	0,6510100	0,3383317	1,2526581



On the basis of the significance level $\alpha = 0.05$, based on the obtained value $p=0.0692$ in the log-rank test ($p=0.09279$ for Gehan's and $p=0.0779$ for Tarone-Ware) we conclude that there is no basis for the rejection of the hypothesis \mathcal{H}_0 . The length of life calculated for the patients in the three compared age groups is similar. However, it is noticeable that the values are quite near to the standard significance level 0.05.

When examining the hazard values (the ratio of the observed values and the expected failure events) we notice that they are a little higher with each age group (0.68, 0.93, 1.43). Although no statistically significant differences among them are seen it is possible that a growth trend of the hazard value (trend in the position of the survival rates) will be found.

Trend for many survival curves

If we introduce into the test the information about the ordering of the compared categories (we will use the age variable in which the age ranges will be numbered, respectively, 1, 2, and 3), we will be able to check if there is a trend in the compared curves. We will study the following hypotheses:

- \mathcal{H}_0 : a lack of a trend in the survival time curves of the patients after a transplantation (a trend dependent on the age of the patients at the time of a transplantation),
- \mathcal{H}_1 : the older the patients at the time of a transplantation, the greater/smaller the probability of their survival over a given period of time.

For trend:	
Chi-square statistic	5,113469574
Degrees of freedom	1
p-value	0,023740797

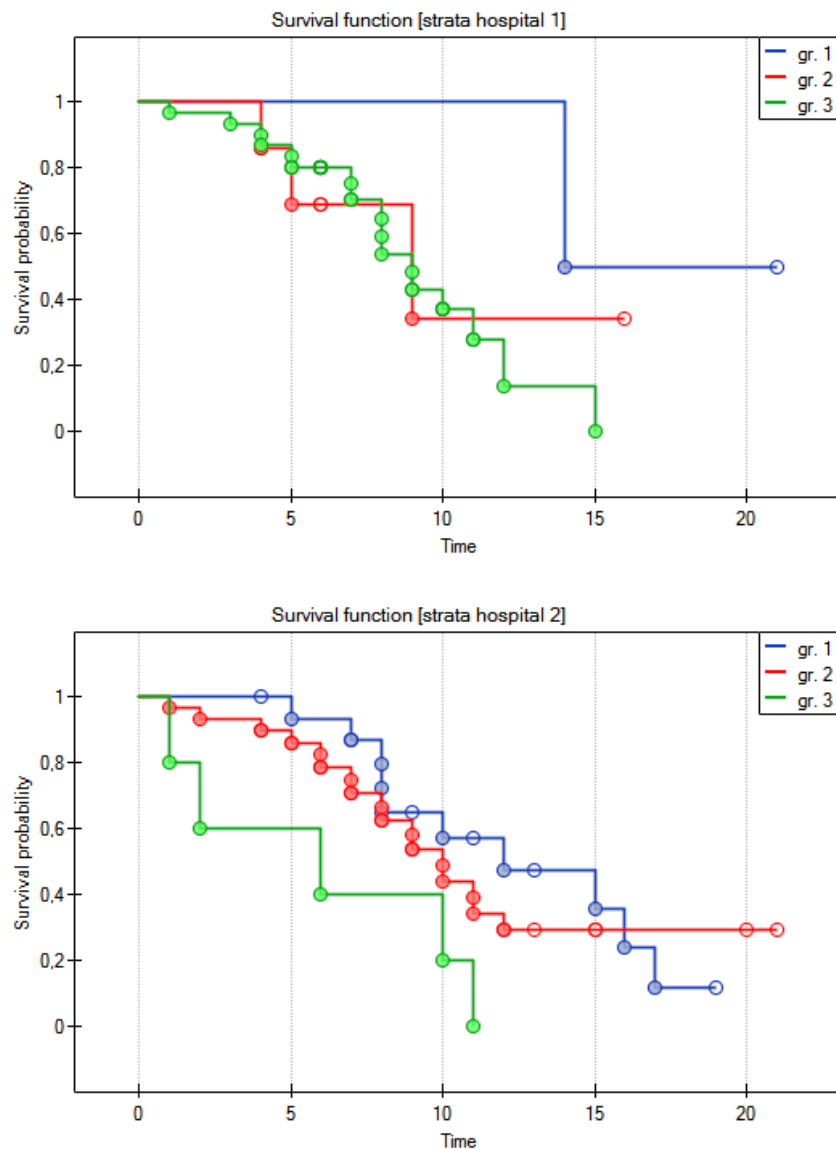
On the basis of the significance level $\alpha = 0.05$, based on the obtained value $p=0.0237$ in the log-rank test ($p=0.0317$ for Gehan's and $p=0.0241$ for Tarone-Ware) we conclude that the survival curves are positioned in a certain trend. On the Kaplan-Meier graph the curve for people aged (55 years; 60 years) is the lowest. Above that curve there is the curve for patients aged (50 years; 55 years). The highest curve is the one for patients aged (45 years; 50 years). Thus, the older the patient at the time of a transplantation, the lower the probability of survival over a certain period of time.

Survival curves for stratas

Let us now check if the trend observed before is independent of the hospital in which the transplantation took place. For that purpose we will choose a hospital as the stratum variable.

Comparison groups	
Analysis time	0,23sec.
Analysed variables	time;status
Significance level	0,05
Grouping variable	age (ordinal)(1;2;3)
Strata variable	hospital
Frequency	89
Failure events	dead
Censored	alive
Test: LogRank	
Strata: hospital 1	
Chi-square statistic	2,413735014
Degrees of freedom	2
p-value	0,299132845
For trend:	
Chi-square statistic	2,357088823
Degrees of freedom	1
p-value	0,124714733
Strata: hospital 2	
Chi-square statistic	5,542657238
Degrees of freedom	2
p-value	0,062578806
For trend:	
Chi-square statistic	3,028300936
Degrees of freedom	1
p-value	0,081823659
Common for stratas	
Chi-square statistic	6,259357261
Degrees of freedom	2
p-value	0,043731849
For trend:	
Chi-square statistic	5,374392333
Degrees of freedom	1
p-value	0,020434458

Logrank				
Strata	Group	Obs.	Exp.	Obs./Exp.
hospital 1	1	1	3,1484541	0,3176161
hospital 1	2	3	3,3935138	0,8840394
hospital 1	3	17	14,458031	1,1758170
hospital 2	1	10	12,506498	0,7995843
hospital 2	2	17	17,490412	0,9719610
hospital 2	3	5	2,0030892	2,4961444
Common	1	11	15,654952	0,7026530
Common	2	20	20,883926	0,9576743
Common	3	22	16,461121	1,3364824



The report contains, firstly, an analysis of the strata: both the test results and the hazard ratio. In the first stratum the growing trend of hazard is visible but not significant. In the second stratum a trend with the same direction (a result bordering on statistical significance) is observed. A cumulation of those trends in a common analysis of strata allowed the obtainment of the significance of the trend of the survival curves. Thus, the older the patient at the time of a transplantation, the lower the probability of survival over a certain period of time, independently from the hospital in which the transplantation took place.

A comparative analysis of the survival curves, corrected by strata, yields a result significant for the log-rank and Tarone-Ware tests and not significant for Gehan's test, which might mean that the differences among the curves are not so visible in the initial survival periods as in the later ones. By looking at the hazard ratio of the curves compared in pairs

Logrank				
Strata	Group	Hazard r.	-95%CI	+95%CI
hospital 1	1<=>2	0,3592782	0,0775124	1,6652928
hospital 1	1<=>3	0,2701238	0,0798335	0,9139877
hospital 1	2<=>3	0,7518511	0,2305103	2,4522984
hospital 2	1<=>2	0,8226505	0,3981102	1,6999158
hospital 2	1<=>3	0,3203277	0,0720764	1,4236263
hospital 2	2<=>3	0,3893849	0,0902500	1,6800069
Common	1<=>2	0,7337077	0,3810310	1,4128167
Common	1<=>3	0,5257480	0,2631977	1,0502028
Common	2<=>3	0,7165633	0,3755802	1,3671193

we can localize significant differences. For the comparison of the curve of the youngest group with the curve of the oldest group the hazard ratio is the smallest, 0.53, the 95% confidence interval for that ratio, (0.26 ; 1.05), does contain value 1 but is on the verge of that value, which can suggest that there are significant differences between the respective curves. In order to confirm that supposition an inquisitive researcher can, with the use of the data filter in the analysis window, compare the curves in pairs.

Data Filter		
variable	condition	value
4-age (nominal)	=	<45; 50)
4-age (nominal)	=	<50; 55)

However, it ought to be remembered that one of the corrections for multiple comparisons should be used and the significance level should be modified. In this case, for Bonferroni's correction, with three comparisons, the significance level will be 0.017. For simplicity, we will only avail ourselves of the log-rank test.

(45 years; 50 years) vs (50 years; 55 years)

Common for stratas	
Chi-square statistic	0,588444784
Degrees of freedom	1
p-value	0,443021142

(45 years; 50 years) vs (55 years; 60 years)

Common for stratas	
Chi-square statistic	8,944721898
Degrees of freedom	1
p-value	0,002782725

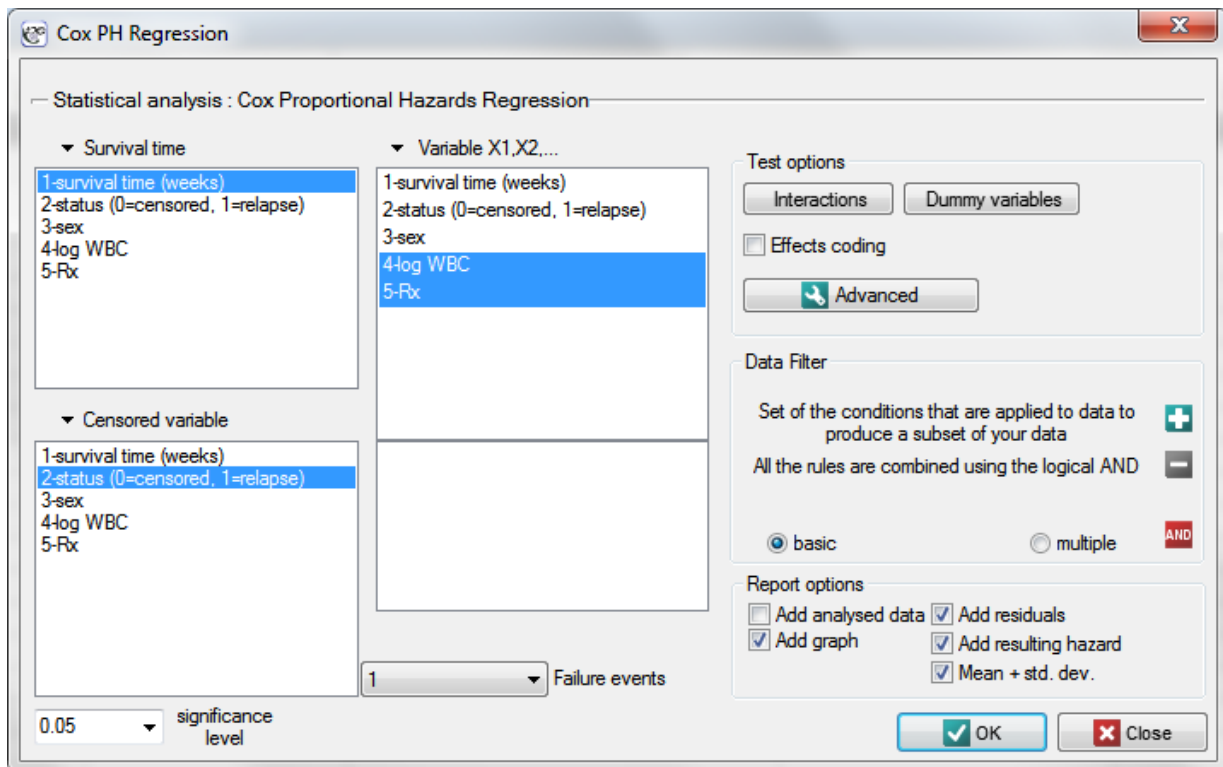
(50 years; 55 years) vs (55 years; 60 years)

Common for stratas	
Chi-square statistic	2,241246614
Degrees of freedom	1
p-value	0,134372609

As expected, statistically significant differences only concern the survival curves of the youngest and oldest groups.

19.4 PROPORTIONAL COX HAZARD REGRESSION

The window with settings for Cox regression is accessed via the menu Statistics→Survival analysis→PH Cox regression



Cox regression, also known as the Cox proportional hazard model, is the most popular regressive method for survival analysis. It allows the study of the impact of many independent variables (X_1, X_2, \dots, X_k) on survival rates. The approach is, in a way, non-parametric, and thus encumbered with few assumptions, which is why it is so popular. The nature or shape of the hazard function does not have to be known and the only condition is the assumption which also pertains to most parametric survival models, i.e. hazard proportionality.

The function on which Cox proportional hazard model is based describes the resulting hazard and is the product of two values only one of which depends on time (t):

$$h(t, X_1, X_2, \dots, X_k) = h_0(t) \cdot \exp \left(\sum_{i=1}^k \beta_i X_i \right),$$

where:

$h(t, X_1, X_2, \dots, X_k)$ –the resulting hazard describing the risk changing in time and dependent on other factors, e.g. the treatment method,

$h_0(t)$ –the baseline hazard, i.e. the hazard with the assumption that all the explanatory variables are equal to zero,

$\sum_{i=1}^k \beta_i X_i$ –a combination (usually linear) of independent variables and model parameters,

X_1, X_2, \dots, X_k –explanatory variables independent of time,

$\beta_1, \beta_2, \dots, \beta_k$ –parameters.

Dummy variables and interactions in the model

A discussion of the coding of dummy variables and interactions is presented in chapter 17.1 Preparation of the variables for the analysis in multidimensional models).

Correction for ties in Cox regression is based on Breslow's method[14]

The model can be transformed into a the linear form:

$$\ln \left(\frac{h(t, X_1, X_2, \dots, X_k)}{h_0(t)} \right) = \sum_{i=1}^k \beta_i X_i.$$

In such a case, the solution of the equation is the vector of the estimates of parameters $\beta_0, \beta_1, \dots, \beta_k$ called **regression coefficients**:

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}.$$

The coefficients are estimated by the so-called **partial maximum likelihood estimation**. The method is called "partial" as the search for the maximum of the likelihood function L (the program makes use of the Newton-Raphson iterative algorithm) only takes place for complete data; censored data are taken into account in the algorithm but not directly.

There is a certain error of estimation for each coefficient. The magnitude of that error is estimated from the following formula:

$$SE_b = \sqrt{\text{diag}(H^{-1})_b}$$

where:

$\text{diag}(H^{-1})$ is the main diagonal of the covariance matrix.

Note!

When building a model it ought to be remembered that the number of observations should be ten times greater than or equal to the ratio of the estimated model parameters (k) and the smaller one of the proportions of the censored or complete sizes (p), i.e. ($n \geq 10k/p$) Peduzzi P., et al(1995)[67].

Note!

When building the model you need remember that the independent variables should not be multicollinear. In a case of multicollinearity estimation can be uncertain and the obtained error values very high. The multicollinear variables should be removed from the model or one independent variable should be built of them, e.g. instead of the multicollinear variables of mother age and father age one can build the parents age variable.

Note!

The criterion of convergence of the function of the Newton-Raphson iterative algorithm can be controlled with the help of two parameters: the limit of convergence iteration (it gives the maximum number of iterations in which the algorithm should reach convergence) and the convergence criterion (it gives the value below which the received improvement of estimation shall be considered to be insignificant and the algorithm will stop).

19.4.1 Hazard ratio

An individual hazard ratio (HR) is now calculated for each independent variable :

$$HR_i = e^{\beta_i}.$$

It expresses the change of the risk of a failure event when the independent variable grows by 1 unit. The result is adjusted to the remaining independent variables in the model — it is assumed that they remain stable while the studied independent variable grows by 1 unit.

The HR value is interpreted as follows:

- $HR > 1$ means the stimulating influence of the studied independent variable on the occurrence of the failure event, i.e. it gives information about how much greater the risk of the occurrence of the failure event is when the independent variable grows by 1 unit.
- $HR < 1$ means the destimulating influence of the studied independent variable on the occurrence of the failure event, i.e. it gives information about how much lower the risk is of the occurrence of the failure event when the independent variable grows by 1 unit.
- $HR \approx 1$ means that the studied independent variable has no influence on the occurrence of the failure event (1).

Note!

If the analysis is made for a model other than linear or if interaction is taken into account, then, just as in the [logistic regression](#) model we can calculate the appropriate HR on the basis of the general formula which is a combination of independent variables.

19.4.2 Model verification

Statistical significance of particular variables in the model (significance of the odds ratio)

On the basis of the coefficient and its error of estimation we can infer if the independent variable for which the coefficient was estimated has a significant effect on the dependent variable. For that purpose we use Wald test.

Hypotheses:

$$\begin{array}{ll} \mathcal{H}_0 : \beta_i = 0, & \text{or, equivalently: } \mathcal{H}_0 : OR_i = 1, \\ \mathcal{H}_1 : \beta_i \neq 0. & \mathcal{H}_1 : OR_i \neq 1. \end{array}$$

The Wald test statistics is calculated according to the formula:

$$\chi^2 = \left(\frac{b_i}{SE_{b_i}} \right)^2$$

The statistic asymptotically (for large sizes) has the χ^2 [distribution](#) with 1 degree of freedom. On the basis of [test statistics](#), p value is estimated and then compared with the significance level α :

$$\begin{array}{ll} \text{if } p \leq \alpha & \implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0. \end{array}$$

The quality of the constructed model

A good model should fulfill two basic conditions: it should fit well and be possibly simple. The quality of Cox proportional hazard model can be evaluated with a few general measures based on: L_{FM} –the maximum value of likelihood function of a full model (with all variables), L_0 –the maximum value of the likelihood function of a model which only contains one free word, d –the observed number of failure events (in models other than Cox's n , i.e. sample size, is used instead of d).

- **Information criteria** are based on the information entropy carried by the model (model insecurity), i.e. they evaluate the lost information when a given model is used to describe the studied phenomenon. We should, then, choose the model with the minimum value of a given information criterion.

AIC , AIC_c , and BIC is a kind of a compromise between the good fit and complexity. The second element of the sum in formulas for information criteria (the so-called penalty function) measures the simplicity of the model. That depends on the number of parameters (k) in the model and the number of complete observations (d). In both cases the element grows with the increase of the number of parameters and the growth is the faster the smaller the number of observations.

The information criterion, however, is not an absolute measure, i.e. if all the compared models do not describe reality well, there is no use looking for a warning in the information criterion.

- Akaike information criterion

$$AIC = -2 \ln L_{FM} + 2k,$$

It is an asymptomatic criterion, appropriate for large sample sizes.

- Corrected Akaike information criterion

$$AIC_c = AIC + \frac{2k(k+1)}{d-k-1},$$

Because the correction of the Akaike information criterion concerns the sample size (the number of failure events) it is the recommended measure (also for smaller sizes).

- Bayesian information criterion or Schwarz criterion

$$BIC = -2 \ln L_{FM} + k \ln(d),$$

Just like the corrected Akaike criterion it takes into account the sample size (the number of failure events), Volinsky and Raftery (2000)[78].

- **Pseudo R^2** –the so-called McFadden R^2 is a goodness of fit measure of the model (an equivalent of the coefficient of multiple determination R^2 defined for multiple linear regression). The value of that coefficient falls within the range of $< 0; 1$, where values close to 1 mean excellent goodness of fit of the model, 0 — a complete lack of fit. Coefficient R_{Pseudo}^2 is calculated according to the formula:

$$R_{Pseudo}^2 = 1 - \frac{\ln L_{FM}}{\ln L_0}.$$

As coefficient R_{Pseudo}^2 does not assume value 1 and is sensitive to the amount of variables in the model, its corrected value is calculated:

$$R_{Nagelkerke}^2 = \frac{1 - e^{-(2/d)(\ln L_{FM} - \ln L_0)}}{1 - e^{(2/d) \ln L_0}} \quad \text{lub} \quad R_{Cox-Snell}^2 = 1 - e^{\frac{(-2 \ln L_0) - (-2 \ln L_{FM})}{d}}.$$

- **Statistical significance of all variables in the model**

The basic tool for the evaluation of the significance of all variables in the model is **the Likelihood Ratio test**. The test verifies the hypothesis:

$$\begin{aligned}\mathcal{H}_0 : & \quad \text{all } \beta_i = 0, \\ \mathcal{H}_1 : & \quad \text{there is } \beta_i \neq 0.\end{aligned}$$

The test statistic has the form presented below:

$$\chi^2 = -2 \ln(L_0/L_{FM}) = -2 \ln(L_0) - (-2 \ln(L_{FM})).$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with k degrees of freedom.

On the basis of **test statistics, p value** is estimated and then compared with α :

$$\begin{aligned}\text{if } p \leq \alpha & \implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha & \implies \text{there is no reason to reject } \mathcal{H}_0.\end{aligned}$$

19.4.3 Analysis of model residuals

The analysis of the of the model residuals allows the verification of its assumptions. The main goal of the analysis in Cox regression is the localization of outliers and the study of hazard proportionality. Typically, in regression models residuals are calculated as the differences of the observed and predicted values of the dependent variable. However, in the case of censored values such a method of determining the residuals is not appropriate. In the program we can analyze residuals described as: Martingale, deviance, and Schoenfeld. The residuals can be drawn with respect to time or independent variables.

Hazard proportionality assumption

A number of graphical methods for evaluating the goodness of fit of the proportional hazard model have been created (Lee and Wang 2003[49]). The most widely used are the methods based on the model residuals. As in the case of other graphical methods of evaluating hazard proportionality this one is a subjective method. For the assumption of proportional hazard to be fulfilled, the residuals should not form any pattern with respect to time but should be randomly distributed around value 0.

Martingale – the residuals can be interpreted as a difference in time $[0, t]$ between the observed number of failure events and their number predicted by the model. The value of the expected residuals is 0 but they have a diagonal distribution which makes it more difficult to interpret the graph (they are in the range of $-\infty$ to 1).

Deviance – similarly to martingale, asymptotically they obtain value 0 but are distributed symmetrically around zero with standard deviation equal to 1 when the model is appropriate. The deviance value is positive when the studied object survives for a shorter period of time than the one expected on the basis of the model, and negative when that period is longer. The analysis of those residuals is used in the study of the proportionality of the hazard but it is mainly a tool for identifying outliers. In the residuals report those of them which are further than 3 standard deviations away from 0 are marked in red.

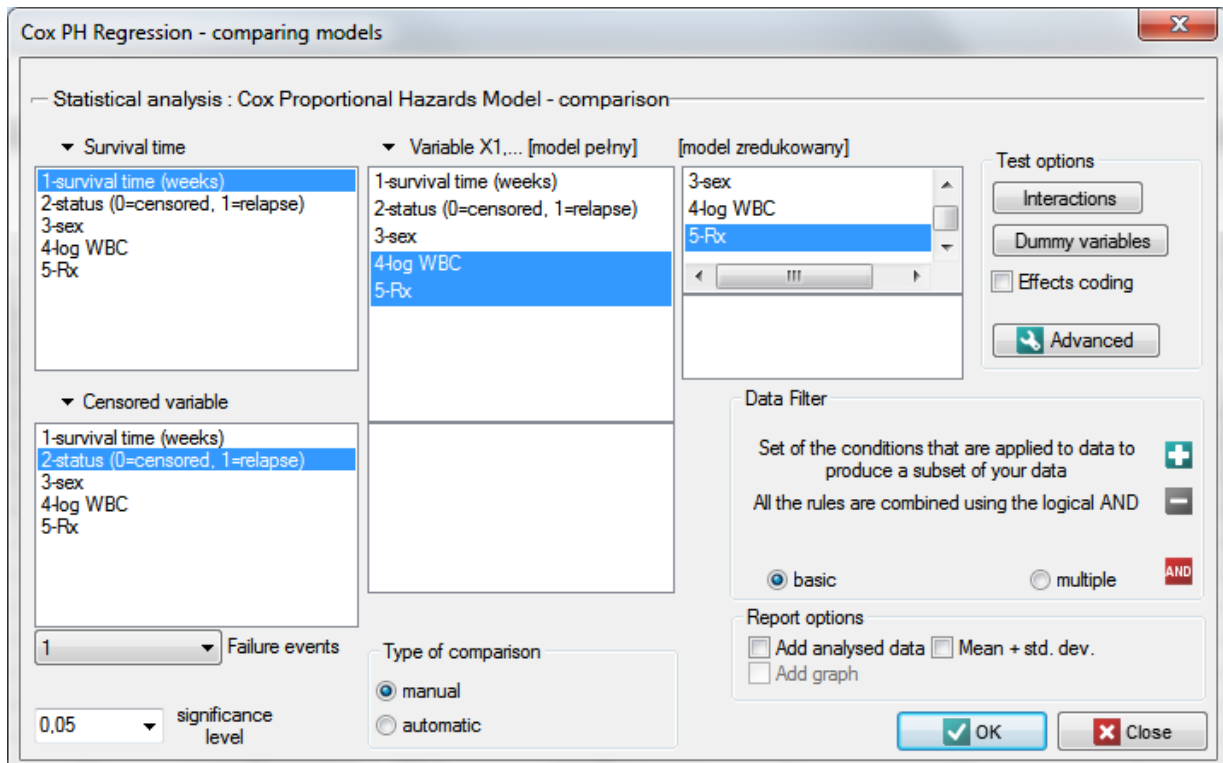
Schoenfeld – the residuals are calculated separately for each independent variable and only defined for complete observations. For each independent variable the sum of Shoenfeld residuals and their expected value is 0. An advantage of presenting the residuals with respect to time for each variable is the possibility of identifying a variable which does not fulfill, in the model, the assumption of hazard proportionality. That is the variable for which the graph of the residuals forms a systematic pattern (usually the studied area is the linear dependence of the residuals on time).

An even distribution of points with respect to value 0 shows the lack of dependence of the residuals on time, i.e. the fulfillment of the assumption of hazard proportionality by a given variable in the model.

If the assumption of hazard proportionality is not fulfilled for any of the variables in Cox model, one possible solution is to make Cox's analyses separately for each level of that variable.

19.5 COMPARISON OF COX PH REGRESSION MODELS

The window with settings for model comparison is accessed via the menu Statistics→Survival analysis→Cox PH Regression – comparing models



Due to the possibility of simultaneous analysis of many independent variables in one Cox regression model, there is a problem of selection of an optimum model. When choosing independent variables one has to remember to put into the model variables strongly correlated with the survival time and weakly correlated with one another.

When comparing models with various numbers of independent variables we pay attention to information criteria (AIC , AIC_c , BIC) and to goodness of fit of the model (R^2_{Pseudo} , $R^2_{Nagelkerke}$, $R^2_{Cox-Snell}$). For each model we also calculate the maximum of likelihood function which we later compare with the use of the Likelihood Ratio test.

Hipotezy:

$$\begin{aligned}\mathcal{H}_0 : L_{FM} &= L_{RM}, \\ \mathcal{H}_1 : L_{FM} &\neq L_{RM},\end{aligned}$$

where:

L_{FM} , L_{RM} – the maximum of likelihood function in compared models (full and reduced).

The test statistic has the form presented below:

$$\chi^2 = -2\ln(L_{RM}/L_{FM}) = -2\ln(L_{RM}) - (-2\ln(L_{FM}))$$

The statistic asymptotically (for large sizes) has the χ^2 distribution with $df = k_{FM} - k_{RM}$ degrees of freedom, where k_{FM} i k_{RM} is the number of estimated parameters in compared models.

On the basis of test statistics, p value is estimated and then compared with α :

$$\begin{aligned} \text{if } p \leq \alpha &\implies \text{we reject } \mathcal{H}_0 \text{ and accept } \mathcal{H}_1, \\ \text{if } p > \alpha &\implies \text{there is no reason to reject } \mathcal{H}_0. \end{aligned}$$

We make the decision about which model to choose on the basis of the size: AIC , AIC_c , BIC , R^2_{Pseudo} , $R^2_{Nagelkerke}$, $R^2_{Cox-Snell}$ and the result of the Likelihood Ratio test which compares the subsequently created (neighboring) models. If the compared models do not differ significantly, we should select the one with a smaller number of variables. This is because a lack of a difference means that the variables present in the full model but absent in the reduced model do not carry significant information. However, if the difference is statistically significant, it means that one of them (the one with the greater number of variables) is significantly better than the other one.

In the program PQStat the comparison of models can be done manually or automatically.

- **Manual** model comparison – construction of 2 models:
 - a full model – a model with a greater number of variables,
 - a reduced model – a model with a smaller number of variables – such a model is created from the full model by removing those variables which are superfluous from the perspective of studying a given phenomenon.

The choice of independent variables in the compared models and, subsequently, the choice of a better model on the basis of the results of the comparison, is made by the researcher.

- **Automatic** model comparison is done in several steps:
 - step 1 Constructing the model with the use of all variables.
 - step 2 Removing one variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.
 - step 3 A comparison of the full and the reduced model.
 - step 4 Removing another variable from the model. The removed variable is the one which, from the statistical point of view, contributes the least information to the current model.
 - step 5 A comparison of the previous and the newly reduced model.
 - ...

In that way numerous, ever smaller models are created. The last model only contains 1 independent variable.

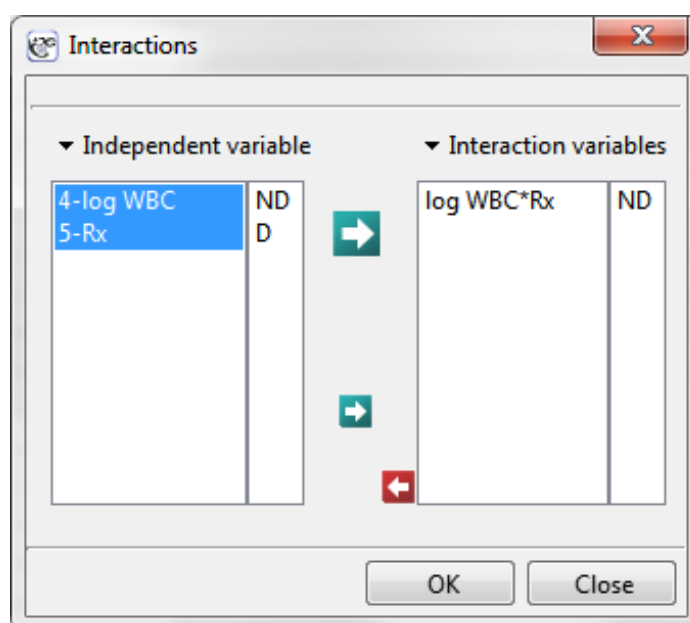
EXAMPLE 19.2. (file: remissionLeukemia.pqs)

The analysis is based on the data about leukemia described in the work of Freirich et al. 1963[32] and further analyzed by many authors, including Kleinbaum and Klein 2005[44]. The data contain information about the time (in weeks) of remission until the moment when a patient was withdrawn from the

study because of an end of remission (a return of the symptoms) or of the censorship of the information about the patient. The end of remission is the result of a failure event and is treated as a **complete** observation. An observation is **censored** if a patient remains in the study to the end and remission does not occur or if the patient leaves the study.

Patients were assigned to one of two groups: a group undergoing treatment (marked as 1) and a placebo group (marked as 0). The information about the patients' sex was gathered (1=man, 0=woman) and about the values of the indicator of the number of white cells, marked as "log WBC", which is a well-known prognostic factor.

The aim of the study is to determine the influence of treatment on the time of remaining in remission, taking into account possible confounding factors and interactions. In the analysis we will focus on the "Rx (1=placebo, 0=treatment)" variable. We will place the "log WBC" variable in the model as a possible confounding factor (which modifies the effect). In order to evaluate the possible interactions of "Rx" and "log WBC" we will also consider a third variable, a ratio of the interacting variables. We will add the variable to the model by selecting, in the analysis window, the Interactions button and by setting appropriate options there.



We build three Cox models:

Model A only contains the "Rx" variable:

Model		B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard rat	-95% CI	+95% CI
	Rx	1.5091913	0.4095644	0.7064599	2.3119228	13.578263	0.0002288	4.5230719	2.0268035	10.093814

Model B contains the "Rx" variable and the potentially confounding variable "log WBC":

Model		B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard rat	-95% CI	+95% CI
	log WBC	1.6043432	0.3293283	0.9588716	2.2498148	23.732115	0.0000011	4.9745910	2.6087511	9.4859783
	Rx	1.2940672	0.4221039	0.4667586	2.1213757	9.3988516	0.0021712	3.6475921	1.5948164	8.3426073

Model C contains the "Rx" variable, the "log WBC" variable, and the potential effect of the interactions of those variables: "Rx × log WBC"

Model	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard rat	-95% CI	+95% CI
log WBC	1.8027879	0.4467169	0.9272388	2.6783371	16.286373	0.0000544	6.0665375	2.5275205	14.560859
Rx	2.3549391	1.6810211	-0.939801	5.6496801	1.9625151	0.1612445	10.537488	0.3907052	284.20054
log WBC*Rx	-0.3421951	0.5197406	-1.3608671	0.6764778	0.4334850	0.5102838			

The variable which informs about the interaction of "Rx" and "log WBC", included in model C, is not significant in model C, according to the Wald test. Thus, we can view further consideration of the interactions of the two variables in the model to be unnecessary. We will obtain similar results by comparing, with the use of a likelihood ratio test, model C with model B. We can make the comparison by choosing the Cox PH regression – comparing models menu. We will then obtain a non-significant result ($p=0.5134$) which means that model C (model with interaction) is NOT significantly better than model B (model without interaction).

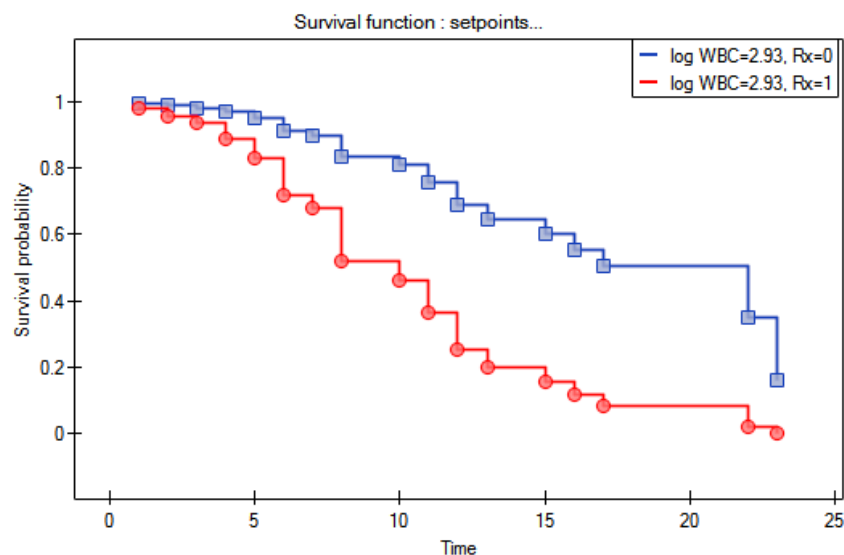
Chi-square - models comparison	0.427079874
Degrees of freedom	1
p-value	0.513425299

Therefore, we reject model C and move to consider model B and model A.

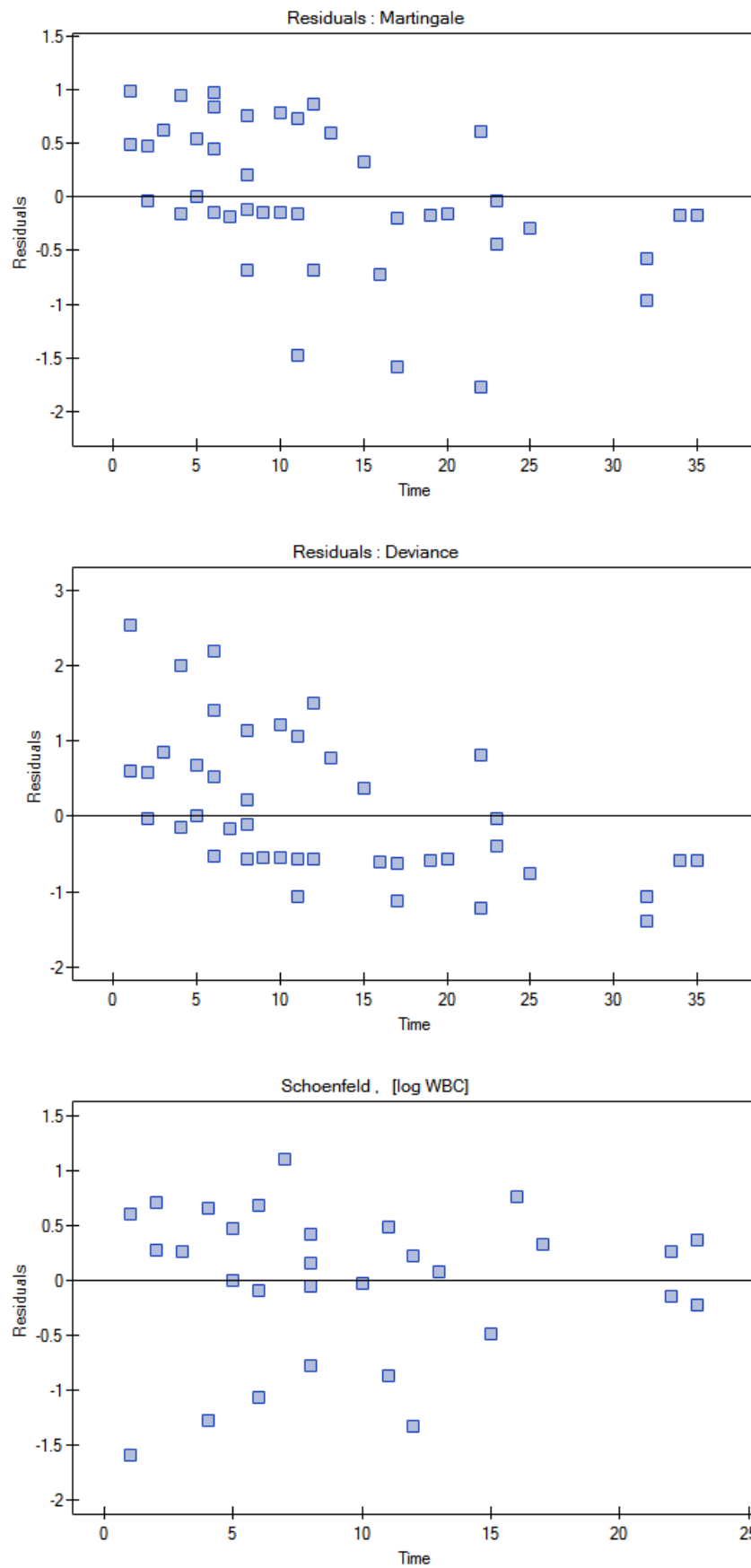
HR for "Rx" in model B is 3.65 which means that hazard for the "placebo group" is about 3.6 greater than for the patients undergoing treatment. Model A only contains the "Rx" variable, which is why it is usually called a "crude" model – it ignores the effect of potential confounding factors. In that model the HR for "Rx" is 4.52 and is much greater than in model B. However, let us look not only at the point values of the HR estimator but also at the 95% confidence interval for those estimators. The range for "Rx" in model A is 8.06 (10.09 minus 2.03) wide and is narrower in model B: 6.74 (8.34 minus 1.60). That is why model B gives a more precise HR estimation than model A. In order to make a final decision about which model (A or B) will be better for the evaluation of the effect of treatment ("Rx") we will once more perform a comparative analysis of the models in the Cox PH regression – comparing models module. This time the likelihood ratio test yields a significant result ($p<0.0001$), which is the final confirmation of the superiority of model B. That model has the lowest value of information criteria ($AIC=148.6$, $AICc=149$, $BIC=151.4$) and high values of goodness of fit ($Pseudo R^2_{McFadden} = 0.2309$, $R^2_{Nagelkerke} = 0.7662$, $R^2_{Cox-Snell} = 0.7647$).

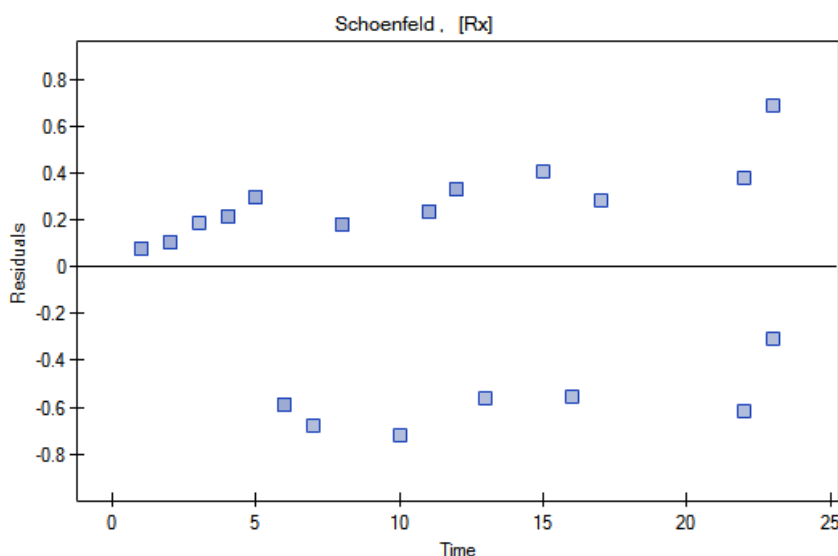
Cox Proportional Hazards Regression - comparison	
Analysis time	0.12sec.
Analysed variables	survival time (weeks);status
Significance level	0.05
Grouping variable	Rx(0;1)
Number of variables in the model 1	2
Convergence criterion met	
-2 Log Likelihood	144.558519789
AIC - Akaike criterion	148.558519789
AICc - corrected Akaike criterion	149.002964233
BIC - Bayesian criterion	151.360914552
Pseudo R2 (McFadden)	0.230949395
R2 (Nagelkerke)	0.766193449
R2 (Coxa-Snella)	0.764737344
Number of variables in the model 2	1
Convergence criterion met	
-2 Log Likelihood	172.759244379
AIC - Akaike criterion	174.759244379
AICc - corrected Akaike criterion	174.902101522
BIC - Bayesian criterion	176.160441761
Pseudo R2 (McFadden)	0.080921681
R2 (Nagelkerke)	0.398474704
R2 (Coxa-Snella)	0.397717426
Chi-square - models comparison	28.20072459
Degrees of freedom	1
p-value	0.000000109

The analysis is complemented with the presentation of the survival curves of both groups, the treatment one and the placebo one, corrected by the influence of "log WBC", for model B. In the graph we observe the differences between the groups, which occur at particular points of survival time. In order to draw such curves, having selected the Add a graph option, we select the Survival function: setpoints... option and set the values for the "Rx" variable as 0 for the first curve (the placebo group) and 1 for the second curve (the treatment group). For the "Log WBC" variable we enter the mean value, i.e. 2.93.



At the end we will evaluate the assumptions of Cox regression by analyzing the model residuals with respect to time.





We do not observe any outliers, however, the martingale and deviance residuals become lower the longer the time. Schoenfeld residuals have a symmetrical distribution with respect to time. In their case the analysis of the graph can be supported with various tests which can evaluate if the points of the residual graph are distributed in a certain pattern, e.g. a linear dependency. In order to make such an analysis we have to copy Schoenfeld residuals, together with time, into a datasheet, and test the type of the dependence which we are looking for. The result of such a test for each variable signifies if the assumption of hazard proportionality by a variable in the model has been fulfilled. It has been fulfilled if the result is statistically insignificant and it has not been fulfilled if the result is statistically significant. As a result the variable which does not fulfill the regression assumption of the Cox proportional hazard can be excluded from the model. In the case of the "Log WBC" and "Rx" variables the symmetrical distribution of the residuals suggests the fulfillment of the assumption of hazard proportionality by those variables. That can be confirmed by checking the correlation, e.g. Pearson's linear or Spearman's monotonic, for those residuals and time.

Later we can add the sex variable to the model. However, we have to act with caution because we know, from various sources, that sex can have an influence on the survival function as regards leukemia, in that survival functions can be distributed disproportionately with respect to each other along the time line. That is why we create the Cox model for three variables: "Sex", "Rx", and "log WBC". Before interpreting the coefficients of the model we will check Schoenfeld residuals. We will present them in graphs and their results, together with time, will be copied from the report to a new data sheet where we will check the occurrence of Spearman's monotonic correlation. The obtained values are $p=0.0259$ (for the time and Schoenfeld residuals correlation for sex), $p=0.6192$ (for the time and Schoenfeld residuals correlation for log WBC), and $p=0.1490$ (for the time and Schoenfeld residuals correlation for Rx) which confirms that the assumption of hazard proportionality has not been fulfilled by the sex variable. Therefore, we will build the Cox models separately for women and men. For that purpose we will make the analysis twice, with the data filter switched on. First, the filter will point to the female sex (0), second, to the male sex (1).

For women

Model	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard rat	-95% CI	+95% CI
log WBC	1.1701250	0.4985684	0.1929488	2.1473011	5.5082668	0.0189267	3.2223954	1.2128207	8.5617207
Rx	0.2667231	0.5659162	-0.842452	1.3758985	0.2221350	0.6374179	1.3056788	0.4306531	3.9586323

For men

Model									
	B coeff.	B error	-95% CI	+95% CI	Wald stat.	p-value	Hazard rat	-95% CI	+95% CI
log WBC	1.6389171	0.5190378	0.6216215	2.6562126	9.9704762	0.0015907	5.1495899	1.8619449	14.242245
Rx	1.8590474	0.7291016	0.4300345	3.2880603	6.5013690	0.0107791	6.4176205	1.5373105	26.790850

20 RELIABILITY ANALYSIS

Reliability analysis is usually associated with the complex scale construction, in particular summary scales (these consist of many individual items). Reliability analysis, associated as its internal consistency, informs us to what extent a particular scale measures what it should measure. In other words, to what extent the scale items measure the things that are measured by the whole scale.

When every scale item measures the same construct (the correlation between the items should be high) we can call it reliable scale. This assumption can be checked by calculating the matrix of [the Pearson's correlation coefficient](#). Many [measures of concordance](#) can be used in reliability analysis. However, the most popular technique is the α -Cronbach coefficient and so-called split-half reliability.

Cronbach's α coefficient was named for the first time in 1951[25], by Cronbach. It measures the proportion of single item variances and the whole scale variance (items sum). It is calculated according to the following formula:

$$\alpha_C = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k sd_i^2}{sd_t^2} \right),$$

where:

k – number of scale items,

sd_i^2 – variance of i item,

sd_t^2 – variance of items sum.

Standardised reliability coefficient $\alpha_{standard}$ is calculated according to the following formula:

$$\alpha_{standard} = \frac{k\bar{r}_p}{1 + (k-1)\bar{r}_p},$$

where:

\bar{r}_p – mean of all the Pearson's correlation coefficients for $(k(k-1)/2)$ scale items.

Alpha can take on any value less than or equal to 1, including negative values, although only positive values make sense. If all scale items are reliable, the reliability coefficient is 1.

There are some values that help in an assessment of particular scale items usefulness:

- the value of α_C coefficient calculated after removing a particular scale item,
- the value of standard deviation of a scale calculated after removing a particular scale item,
- mean value of a scale calculated after removing a particular scale item,
- the Pearson's correlation coefficients between a particular item and the sum of other items.

Split-half reliability

Split-half reliability is a random scale item division into 2 halves and an analysis of the halves correlation. It is carried out by the Spearman-Brown split-half reliability coefficient, published independently by Spearman (1910)[75] and Brown (1910)[17]:

$$r_{SH} = \frac{2r_p^*}{1 + r_p^*},$$

where:

r_p^* – the Pearson's correlation coefficient between halves of a scale.

If two halves, randomly selected, are ideally correlated: $r_{SH} = 1$.

A formula for the split-half reliability coefficient proposed by Guttman (1945)[36]:

$$r_{SHG} = 2 \left(1 - \frac{sd_{t1}^2 + sd_{t2}^2}{sd_t^2} \right),$$

where:

sd_{t1}^2, sd_{t2}^2 – variance of the first and the second half of a scale,

sd_t^2 – variance of the sum of all scales items.

Note

The scale is reliable if the scales reliability coefficients ($\alpha_C, \alpha_{standard}, r_{SH}, r_{SHG}$) are larger than 0.6 and smaller than 1.

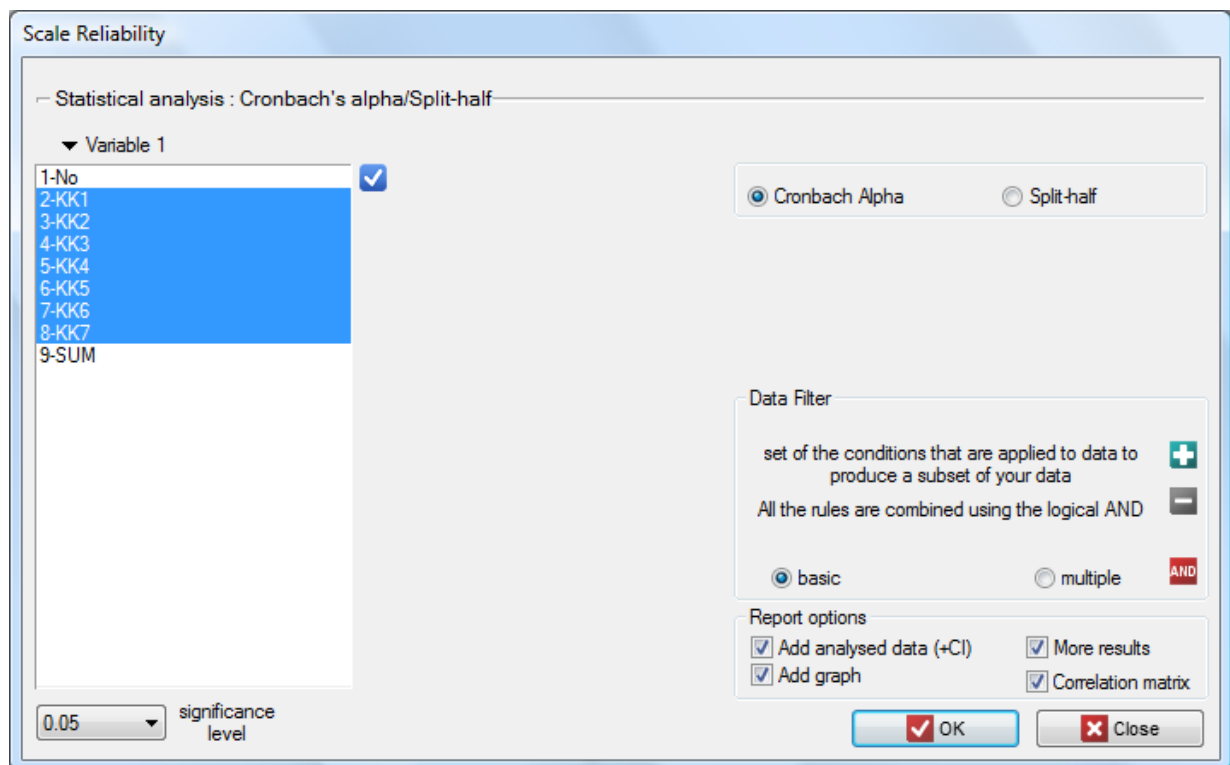
Standard error of measurement is calculated for the reliable scale, according to the following formula:

$$SEM = sd_t \sqrt{1 - \alpha_C} \quad \text{— for the Cronbach's alpha coefficient of reliability}$$

or

$$SEM = sd_t \sqrt{1 - r_{SH}} \quad \text{— for the split-half reliability coefficient}$$

The settings window with the Cronbach's alpha/Split-half can be opened in Statistics menu → Scale reliability.



EXAMPLE 20.1. (scale.pqs file)

A "competence scale", created in some company, enables an assessment of the usefulness of future employees. Apart from participation in a job interview, candidates fill in the questionnaire that includes the "competence scale" questions. There are 7 questions in the scale. For each question, one can get 1 - 5 points, where 1 - the lowest mark, 5 - the highest mark. The maximum score of the questionnaire is 35. In the table, there are scores obtained by 24 candidates.

Lp	KK1	KK2	KK3	KK4	KK5	KK6	KK7	SUMA
1	3	3	5	5	5	5	1	27
2	5	4	4	3	3	5	1	25
3	5	5	3	5	3	2	1	24
4	1	2	5	5	5	5	2	25
5	4	5	5	5	5	5	1	30
6	4	4	5	5	5	5	3	31
7	1	1	5	5	5	5	2	24
8	5	5	5	5	3	5	3	31
9	3	2	2	5	4	2	1	19
10	3	4	3	4	4	2	1	21
11	4	4	3	4	4	4	4	27
12	1	1	3	4	1	1	3	16
13	3	3	4	5	5	5	1	26
14	4	5	5	5	5	5	2	31
15	1	4	4	4	1	4	4	22
16	1	4	5	5	5	5	1	26
17	5	5	5	5	5	5	2	32
18	5	3	5	5	3	5	4	30
19	1	1	2	2	2	1	4	13
20	5	5	5	5	5	5	5	35
21	5	3	5	5	5	5	1	29
22	5	5	5	5	5	1	5	31
23	2	1	5	3	2	4	1	18
24	5	5	5	5	5	5	5	35

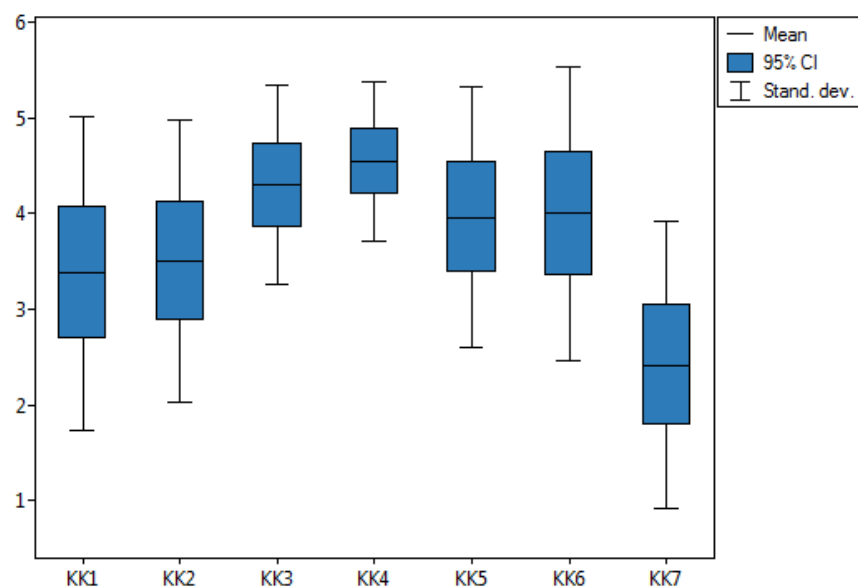
For checking the accuracy of the "competence scale", the reliability should be analysed.

The correlation matrix indicates that the last item is least correlated with the other items. Thus, it is suspected that the item does not measure the same construct as the others.

Correlatio	KK2	KK3	KK4	KK5	KK6	KK7
KK1	0.712	0.265	0.355	0.338	0.225	0.164
KK2	.	0.326	0.443	0.378	0.269	0.196
KK3	.	.	0.512	0.498	0.735	0.058
KK4	.	.	.	0.67	0.409	-0.049
KK5	0.478	-0.14
KK6	-0.151

The competence scale turned out to be a reliable scale. Cronbach alpha coefficient is 0.736805, and mean of all the Pearson's correlation coefficients is 0.31847.

Deleted item	KK7
Scale mean if item deleted	23.666667
Scale standard deviation if item deleted	5.73067
Correlation between deleted item and sum of remaining	0.026954
Cronbach Alpha if item deleted	0.803619
Group size	24
Number of items	7
Mean of scale	26.083333
Standard deviation of scale	5.96305
Cronbach Alpha for scale	0.736805
Standard error of measurement	3.059199
Average correlation between pairs of items	0.31847
Standardized Cronbach alpha	0.765863



A more precised analysis of each item indicates that, except the last one, they all influence scale reliability in a similar way. Correlation between the KK7 item and the other scales items, is the weakest: 0.026954. Removing the KK7 item from the scale, the Cronbach alpha coefficient would increase to 0.803619.

Similar conclusion can be drawn on the basis of split-half reliability analysis, carried out on the items randomly divided into 2 halves (KK1, KK3, KK5) (KK2, KK4, KK6, KK7).

Cronbach's alpha/Split-half	
Analysis time	0.04sec.
Analysed variables	KK1, KK3, KK5, KK2, KK4, KK6, KK7
Significance level	0.05
Group size	24
Mean of scale	26.083333
Standard deviation of scale	5.96305
Correlation between two halves of scale	0.750862
Split-half reliability	0.857705
Standard error of measurement	2.24938
Guttman split-half reliability	0.856531
First half	
Number of items	3
Names of items	KK1, KK3, KK5
Mean	11.625
Standard deviation	3.076029
Cronbach Alpha	0.607122
Second half	
Number of items	4
Names of items	KK2, KK4, KK6, KK7
Mean	14.458333
Standard deviation	3.296628
Cronbach Alpha	0.416958


Spearman-Brown split-half reliability Coefficient is 0.857705. Guttman split-half reliability coefficient is 0.856531. The halves are well correlated – the correlation coefficient is 0.750862. However, the value of Cronbach alpha coefficient is too low for the second half (0.416958). This half includes the KK7 item, which shows a weak correlation with the other scale items. Removing the item and repeating the analysis, all the items are really high and reliable.

Cronbach's alpha/Split-half	
Analysis time	0.03sec.
Analysed variables	KK1, KK3, KK5, KK2, KK4, KK6
Significance level	0.05
Group size	24
Mean of scale	23.666667
Standard deviation of scale	5.73067
Correlation between two halves of scale	0.822933
Split-half reliability	0.902867
Standard error of measurement	1.786032
Guttman split-half reliability	0.902251
First half	
Number of items	3
Names of items	KK1, KK3, KK5
Mean	11.625
Standard deviation	3.076029
Cronbach Alpha	0.607122
Second half	
Number of items	3
Names of items	KK2, KK4, KK6
Mean	12.041667
Standard deviation	2.92633
Cronbach Alpha	0.586418

21 THE WIZARD

The Wizard is a tool which makes the navigation easier to go, through the basic statistics included in an application, especially for a novice user. It includes suggestions of assumptions which should be checked before the choice of a particular [statistic test](#). The last step of the wizard is to select an appropriate statistic test and to open the window with the settings of the test options.

The Wizard may be launched by:

- Statistics→Wizard,
-  button on a toolbar.

A launched wizard window includes the possibility to choose the kind of an analysis that a user wants to carry out. A user may choose:

Comparison – 1 group - to compare values of measurements coming from a 1 population with the specific value given by the user. This population is represented by raw data gathered in a 1 column or cumulated to the form of a frequency table.

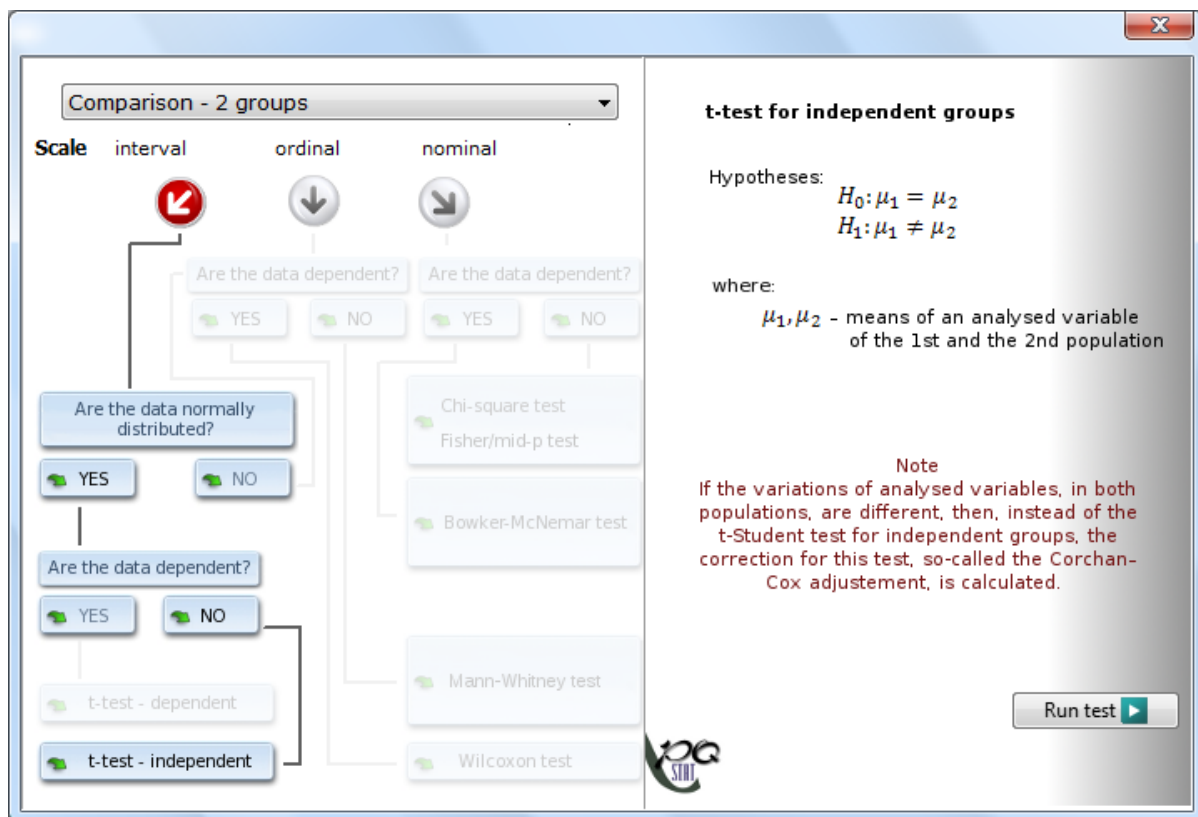
Comparison – 2 groups - to compare values of measurements coming from 2 populations. These populations are represented by raw data gathered in 2 columns or cumulated to the form of a contingency table.

Comparison – more than 2 groups - to compare values of measurements coming from several populations. The populations are represented by data collected in the form of raw data, in several columns.

Correlation - to check the occurrence of dependence between 2 parameters coming from a 1 population. These features are represented by raw data gathered in 2 columns or cumulated to the form of a contingency table.

Agreement - to check the concordance of obtained measurements. These features are represented by raw data gathered in several columns or cumulated to the form of a contingency table.

When the user chooses the kind of an analysis, a graph will occur. The graph is divided according to a scale, on which the measurement of the analysed features was done ([interval scale](#), [ordinal scale](#), [nominal scale](#)).



The user moves on the graph by selecting the adequate answers to the asked questions. After the user gets through the way on the graph, chosen by himself, he is able to perform this test, which – according to the replies – is an appropriate one to solve the determined statistical problem.

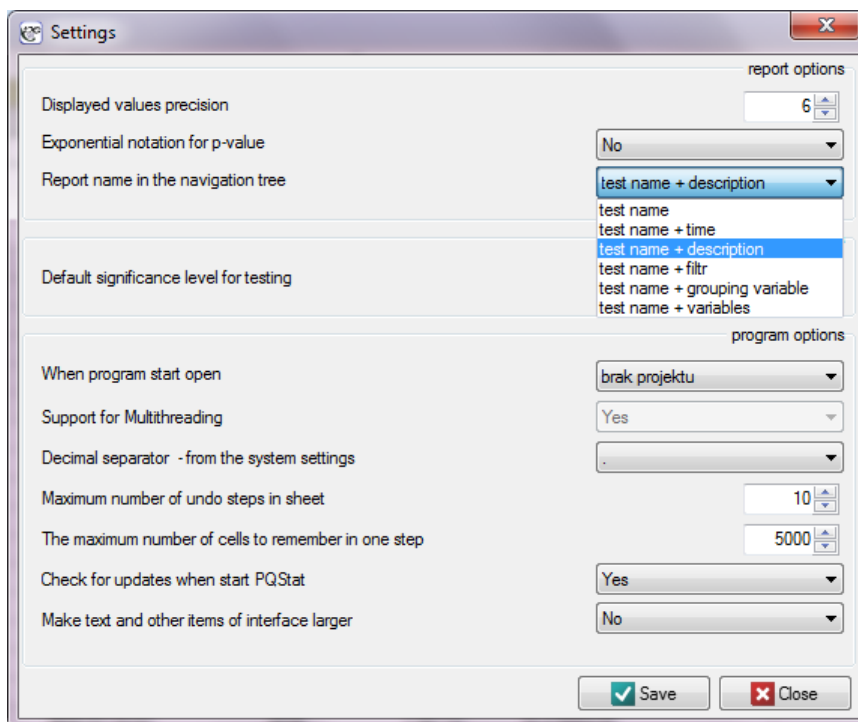
22 OTHER NOTES

22.1 FILES FORMAT

PQS - default file format for PQStat files; is used for representing all objects created with PQStat (project,datasheet,report,graph);

PQX - XML file for PQStat, is used for representing all objects created with PQStat; PQX files are stored in Unicode text format (support UTF-8 character encoding); recommended for use on computers with a small amount of memory.

22.2 SETTINGS



References

- [1] Abdi H. (2007), *Bonferroni and Sidak corrections for multiple comparisons*, in N.J. Salkind (ed.): *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage
- [2] Agresti A., Coull B.A. (1998), *Approximate is better than "exact" for interval estimation of binomial proportions*. *American Statistics* 52: 119-126
- [3] Altman D.G., Bland J.M. (1983), *Measurement in medicine: the analysis of method comparison studies*. *The Statistician* 32: 307–317
- [4] Anscombe F.J. (1981), *Computing in Statistical Science through APL*. Springer-Verlag, New York
- [5] Armitage P., Berry G., (1994), *Statistical Methods in Medical Research* (3rd edition); Blackwell
- [6] Barnard G.A. (1989), *On alleged gains in power from lower p-values*. *Statistics in Medicine* 8:1469-1477
- [7] Beal S.L. (1987), *Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples*. *Biometrics* 43: 941-950.
- [8] Bender R. (2001), *Calculating confidence intervals for the number needed to treat*. *Controlled Clinical Trials* 22:102–110.
- [9] Betty R. Kirkwood and Jonathan A. C. Sterne (2003), *Medical Statistics* (2nd ed.). Meassachusetts: Blackwell Science, 177–188, 240–248
- [10] Bland J.M., Altman D.G. (1986), *Statistical methods for assessing agreement between two methods of clinical measurement*. *Lancet* 327 (8476): 307–10
- [11] Bowker A.H. (1948), *Test for symmetry in contingency tables*. *Journal of the American Statistical Association*, 43, 572-574
- [12] Breslow N.E., Day N.E. (1980), *Statistical Methods in Cancer Research: Vol. I - The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer
- [13] Breslow N.E. (1996), *Statistics in epidemiology: the case-control study*, *Journal of the American Statistical Association*, 91, 14–28
- [14] Breslow N.E. (1974), *Covariance analysis of censored survival data*. *Biometrics*, 30(1):89–99.
- [15] Brown L.D., Cai T.T., DasGupta A. (2001), *Interval Estimation for a Binomial Proportion*. *Statistical Science*, Vol. 16, no. 2, 101-133
- [16] Brown M.B., Forsythe A. B. (1974a), *Robust tests for equality of variances*. *Journal of the American Statistical Association*, 69,364-367
- [17] Brown W. (1910), *Some experimental results in the correlation of mental abilities*. *British Journal of Psychology*, 3, 296-322.
- [18] Clopper C. and Pearson S. (1934), *The use of confidence or fiducial limits illustrated in the case of the binomial*. *Biometrika* 26: 404-413
- [19] Cochran W.G. (1950), *The comparison of percentages in matched samples*. *Biometrika*, 37, 256-266.
- [20] Cochran W.G. (1952), *The chi-square goodness-of-fit test*. *Annals of Mathematical Statistics*, 23,3 15-345,

- [21] Cochran W.G. and Cox G.M. (1957), *Experimental designs (2nd 4.)*. New York: John Wiley and Sons.
- [22] Cohen J. (1960), *A coefficient of agreement for nominal scales*. *Educational and Psychological Measurement*, 10,3746
- [23] Cox D.R. (1972), *Regression models and life tables*. *Journal of the Royal Statistical Society*, B34:187-220.
- [24] Cramkr H. (1946), *Mathematical models of statistics*. Princeton, NJ: Princeton University Press.
- [25] Cronbach L.J. (1951), *Coefficient alpha and the internal structure of tests*. *Psychometrika*, 16(3), 297-334.
- [26] DeLong E.R., DeLong D.M., Clarke-Pearson D.L., (1988), *Comparing the areas under two or more correlated receiver operating curves: A nonparametric approach*. *Biometrics* 44:837-845.
- [27] Fisher R.A. (1934), *Statistical methods for research workers (5th ed.)*. Edinburgh: Oliver and Boyd.
- [28] Fisher R.A. (1935), *The logic of inductive inference*. *Journal of the Royal Statistical Society, Series A*, 98,39-54
- [29] Fisher R.A. (1936), *The use of multiple measurements in taxonomic problems*. *Annals of Eugenics* 7 (2): 179–188
- [30] Fleiss J.L. (1981), *Statistical methods for rates and proportions*. 2nd ed. (New York: John Wiley) 38-46
- [31] Freeman G.H. and Halton J.H. (1951), *Note on an exact treatment of contingency, goodness of fit and other problems of significance*. *Biometrika* 38:141-149
- [32] Freireich E.O., Gehan E., Frei E., Schroeder L.R., Wolman I.J., et al., (1963) *The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia*. *Blood*, 21: 699–716
- [33] Friedman M. (1937), *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*. *Journal of the American Statistical Association*, 32,675-701.
- [34] Gehan E. A. (1965a), *A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples*. *Biometrika*, 52:203—223.
- [35] Gehan E. A. (1965b), *A Generalized Two-Sample Wilcoxon Test for Doubly-Censored Data*. *Biometrika*, 52:650—653.
- [36] Guttman L. (1945), *A basic for analyzing test-retest reliabilit*. *Psychometrika*, 10, 255-282.
- [37] Hanley J.A. (1987), *Standard error of the Kappa statistic*. *Psychological Bulletin*, Vol 102, No. 2, 315 - 321
- [38] Hanley J.A. i Hajian-Tilaki K.O. (1997), *Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update*. *Academic radiology* 4(1):49-58.
- [39] Hanley J.A. i McNeil M.D. (1982), *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. *Radiology* 143(1):29-36.
- [40] Hanley J.A. i McNeil M.D. (1983), *A method of comparing the areas under receiver operating characteristic curves derived from the same cases*. *Radiology* 148: 839-843.

- [41] Kaplan E.L., Meier P. (1958), *Nonparametric estimation from incomplete observations*. *Journal of the American Statistical Association*, 53:457-481.
- [42] Kendall M.G. (1938), *A new measure of rank correlation*. *Biometrika*, 30, 81-93.
- [43] Kendall M.G., Babington-Smith B. (1939), *The problem of m rankings*. *Annals of Mathematical Statistics*, 10, 275-287.
- [44] Kleinbaum D. G., Klein M., (2005) *Survival Analysis: A Self-Learning Text, Second Edition (Statistics for Biology and Health)*
- [45] Kolmogorov A.N. (1933), *Sulla determinazione empirica di una legge di distribuzione*. *Rivista dell'Inst. Ital. degli. Art.*, 4, 89-91
- [46] Kruskal W.H. (1952), *A nonparametric test for the several sample problem*. *Annals of Mathematical Statistics*, 23, 525-540
- [47] Kruskal W.H., Wallis W.A. (1952), *Use of ranks in one-criterion variance analysis*. *Journal of the American Statistical Association*, 47, 583-621
- [48] Lancaster H.O. (1961), *Significance tests in discrete distributions*. *Journal of the American Statistical Association* 56:223-234
- [49] Lee E. T., Wang J. W. (2003), *Statistical Methods for Survival Data Analysis (ed. third, Wiley 2003)*
- [50] Levene H. (1960), *Robust tests for the equality of variance*. In I. Olkin (Ed.) *Contributions to probability and statistics* (278-292). Palo Alto, CA: Stanford University Press
- [51] Lilliefors H.W. (1967), *On the Kolmogorov-Smirnov test for normality with mean and variance unknown*. *Journal of the American Statistical Association*, 62,399-402
- [52] Lilliefors H.W. (1969), *On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown*. *Journal of the American Statistical Association*, 64,387-389
- [53] Lilliefors H.W. (1973), *The Kolmogorov-Smirnov and other distance tests for the gamma distribution and for the extreme-value distribution when parameters must be estimated*. Department of Statistics, George Washington University, unpublished manuscript
- [54] Lund R.E., Lund J.R. (1983), *Algorithm AS 190, Probabilities and Upper Quantiles for the Studentized Range*. *Applied Statistics*; 34
- [55] Mann H. and Whitney D. (1947), *On a test of whether one of two random variables is stochastically larger than the other*. *Annals of Mathematical Statistics*, 18, 504
- [56] Mantel N. and Haenszel W. (1959), *Statistical aspects of the analysis of data from retrospective studies of disease*. *Journal of the National Cancer Institute*, 22,719-748.
- [57] Mantel N. (1963), *Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure*. *J. Am. Statist. Assoc.*, 58, 690-700.
- [58] Mantel N. (1966), *Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration*. *Cancer Chemotherapy Reports*, 50:163—170.
- [59] Marascuilo L.A. and McSweeney M. (1977), *Nonparametric and distribution-free method for the social sciences*. Monterey, CA: Brooks/Cole Publishing Company

- [60] Marascuilo L.A. and McSweeney M. (1977), *Nonparametric and distribution-free method for the social sciences*. Monterey, CA: Brooks/Cole Publishing Company
- [61] McNemar Q. (1947), *Note on the sampling error of the difference between correlated proportions or percentages*. *Psychometrika*, 12, 153-157
- [62] Mehta C.R. and Patel N.R. (1986), *Algorithm 643. FEXACT: A Fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables*. *ACM Transactions on Mathematical Software*, 12, 154-161
- [63] Miettinen O.S. (1985), *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. John Wiley and Sons, New York
- [64] Miettinen O.S. and Nurminen M. (1985), *Comparative analysis of two rates*. *Statistics in Medicine* 4: 213-226
- [65] Newcombe R.G. (1998), *Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods*. *Statistics in Medicine* 17: 873-890.
- [66] Newman S.C.(2001), *Biostatistical Methods in Epidemiology*. 2nd ed. (New York: John Wiley)
- [67] Peduzzi P., Concato J., Feinstein A.R., Holford T.R. (1995), *Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates*. *Journal of Clinical Epidemiology*, 48:1503-1510
- [68] Plackett R.L. (1984), *Discussion of Yates' "Tests of significance for 2x2 contingency tables"*. *Journal of Royal Statistical Society Series A* 147:426-463
- [69] Pratt J.W. and Gibbons J.D. (1981), *Concepts of Nonparametric Theory*. Springer-Verlag, New York
- [70] Robins, J., Breslow, N., and Greenland S. (1986), *Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models*. *Biometrics* 42, 311-323
- [71] Robins, J., Greenland S. and Breslow, N.E. (1986), *A general estimator for the variance of the Mantel-Haenszel odds ratio*. *American Journal of Epidemiology* 124, 719-723
- [72] Rothman K.J., Greenland S., Lash T.L. (2008), *Modern Epidemiology*, 3rd ed. (Lippincott Williams and Wilkins) 221-225
- [73] Satterthwaite F.E. (1946), *An approximate distribution of estimates of variance components*. *Biometrics Bulletin*, 2, 1 10-1 14
- [74] Savin N.E. and White K.J. (1977), *The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors*. *Econometrica* 45, 1989-1996.
- [75] Spearman C. (1910), *Correlation calculated from faulty data*. *British Journal of Psychology*, 3, 271-295.
- [76] Tarone R. E., Ware J. (1977), *On distribution-free tests for equality of survival distributions*. *Biometrika*, 64(1):156-160.
- [77] Tarone R.E. (1985), *On heterogeneity tests based on efficient scores*. *Biometrika* 72, 91-95
- [78] Volinsky C.T., Raftery A.E. (2000) , *Bayesian information criterion for censored survival models*. *Biometrics*, 56(1):256-262.

- [79] Wallenstein S. (1997), *A non-iterative accurate asymptotic confidence interval for the difference between two Proportions*. *Statistics in Medicine* 16: 1329-1336
- [80] Wallis W.A. (1939), *The correlation ratio for ranked data*. *Journal of the American Statistical Association*, 34,533-538.
- [81] Wilcoxon F. (1945), *Individual comparisons by ranking methods*. *Biometries*, 1,80-83
- [82] Wilcoxon F. (1945), *Individual comparisons by ranking methods*. *Biometries*, 1,80-83
- [83] Wilcoxon F. (1949), *Some rapid approximate statistical procedures*. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation
- [84] Wilcoxon F. (1949), *Some rapid approximate statistical procedures*. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation
- [85] Wilcoxon F. (1949), *Some rapid approximate statistical procedures*. Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation
- [86] Wilson E.B. (1927), *Probable Inference, the Law of Succession, and Statistical Inference*. *Journal of the American Statistical Association*: 22(158):209-212.
- [87] Yates F. (1934), *Contingency tables involving small numbers and the chi-square test*. *Journal of the Royal Statistical Society*, 1,2 17-235
- [88] Yule G. (1900), *On the association of the attributes in statistics: With illustrations from the material of the childhood society, and c*. *Philosophical Transactions of the Royal Society, Series A*, 194,257-319
- [89] Zweig M.H., Campbell G. (1993), *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine*. *Clinical Chemistry* 39:561-577.